The choice of group norms under indirect reciprocity

Marcel Lumkowsky*

October 16, 2025

Preliminary version – Please do not cite or distribute

Abstract

Indirect reciprocity—cooperation maintained through reputation rather than repeated interaction—relies on shared norms for moral evaluation. This experiment examines whether individuals prefer norms based solely on observed actions or those incorporating contextual information, such as a partner's reputation. In a simple game, participants make cooperation decisions in one-shot interactions using second-order information: how a potential recipient behaved previously and the reputational context. They report personal beliefs about appropriate behaviors and perceptions of others' beliefs. A coordination stage allows participants to indicate group-appropriate behaviors, with group norms either revealed or withheld. The 2 × 2 design varies cooperation cost (*Low* vs. *High*) and group norm information (*NoInfo* vs. *Info*) to examine effects on behavioral convergence and cooperation. Results show participants personally favor action-based norms and rarely endorse context-sensitive norms, while group-level assessments tend to differentiate justified from unjustified defections. Providing group norm information modestly increases cooperation under low-cost conditions but does not improve alignment with the group norm.

Keywords: indirect reciprocity, reputation, cooperation, strategic uncertainty, social dilemmas

JEL Codes: C72, C73, C91

*Institute of Economics, University of Kassel, 34109 Kassel, Germany; Email: marcel.lumkowsky@uni-kassel.de

1 Introduction

Imagine Alice on a crowded bus. Another passenger, Bob, asks her to give up her seat. Alice remembers seeing Bob in a similar situation a few weeks earlier. Back then, Charlie asked Bob for his seat, and Bob refused. Remembering this, Alice views Bob as undeserving of help and remains seated. Later, Bob asks Dave for his seat. Like Alice, Dave recalls Bob's earlier refusal. However, Dave does not focus solely on that; he also takes into account that Charlie is notorious for never offering his seat to anyone. With this context, Dave interprets Bob's refusal not as selfish, but as justified—and still considers Bob worthy of help. Indeed, had Dave known that Bob did give up his seat to Charlie, he might have judged Bob negatively, seeing that act as enabling antisocial behavior.

This simple story highlights fundamental features of the cooperation mechanism explored in this paper: indirect reciprocity (Alexander 1987; Nowak 2006; Rand and Nowak 2013). Unlike direct reciprocity, which relies on repeated interactions between the same individuals to reward cooperation or punish defection (Axelrod 2006; Trivers 1971), indirect reciprocity operates through reputation. Individuals observe reputations based on how others treat third parties and adjust their behavior accordingly. Thus, Alice and Dave decide whether to help Bob not based on how Bob treated them personally, but on how he treated someone else—his reputation. However, there is no single, objective standard for updating reputations. Even in this simple example, it is clear that translating actions into reputational judgments is complex. Should Bob's reputation be assessed solely on his behavior, as Alice does, or should the broader context—including Charlie's reputation—also be taken into account, as Dave does? Moreover, if context matters, does it influence only the judgment of negative actions (e.g., refusing to help), or should it also shape how positive actions (e.g., helping) are evaluated? In the theoretical literature on indirect reciprocity, these varying conceptions of what constitutes good behavior are known as assessment rules or norms.[1] The experiment

[1] In the theoretical literature, the terms norm and assessment rule are often used interchangeably to describe principles that

presented in this paper focuses explicitly on these different norms by investigating which types of norms participants find personally appropriate, how they believe others view those norms, and whether groups can coordinate on shared standards for assessing reputations.

This study builds on prior experimental research on indirect reciprocity, which can be broadly categorized into two strands: those that focus on the choice and application of different norms (e.g., Bolton et al. 2005; Gaudeul et al. 2021; Milinski et al. 2001; Swakman et al. 2016), and those that do not (e.g., Engelmann and Fischbacher 2009; Seinen and Schram 2006; Wedekind and Milinski 2000). These strands primarily differ in the type of reputational information presented to participants. Since the present study explicitly investigates how individuals select and apply norms, it aligns most closely with the first group. Among these, the design of the present experiment most closely resembles that of Gaudeul et al. (2021). Like much of the experimental literature on indirect reciprocity, their study employs the "Helping Game" as its central framework. This is a simple, asymmetric interaction in which one player (the donor) must decide whether to *Give* (cooperate) to another player (the recipient) or to *Keep* (defect). *Give* involves transferring a benefit, b, to the recipient at a personal cost, c, to the donor, where $b > c > 0$. If the donor chooses to play *Keep*, neither costs nor benefits are generated. Reputation in this context is based on second-order information, meaning that both the action and some context of the action are revealed. Before making a decision, a donor observes how their current recipient behaved in a previous round when they themselves were in the donor role, as well as what information that recipient had at the time—specifically, whether their own recipient had previously given or kept. This structure produces four distinct reputational signals: (1) *Give* to someone who chose *Give* before; (2) *Keep* to someone who chose *Give* before; (3) *Give* to someone who chose *Keep* before; and (4) *Keep* to someone who chose *Keep* before. In Gaudeul et al. (2021), participants in certain rounds respond using the strategy method—that is, they indicate whether they would play *Give* or *Keep* for

translate observed behavior into reputational standing. In this paper, I use assessment rule to refer to theoretical constructs from evolutionary biology, and norm to refer to the concept as implemented in the experimental context of this study. This distinction is discussed in detail in Section 4.

each of the four possible signals. The resulting pattern of responses reveals the participant's norm. Their findings suggest that most participants either give unconditionally or adopt a norm of giving to those who previously gave, and keeping otherwise—that is, a norm that does not explicitly account for context. Moreover, although responses tend to cluster around four theoretically predicted norms, there remains disagreement among participants about how to translate observed behavior into reputational judgments. Together, these empirical findings—the tendency to judge reputation based solely on actions, and the lack of a unified norm across subjects—suggest that participants do not fully realize the cooperative potential that indirect reciprocity can support. Theoretical models show that assessment rules which explicitly distinguish between justified and unjustified behavior are more effective at sustaining high levels of cooperation (e.g., Leimar and Hammerstein 2001; Ohtsuki and Iwasa 2004). Moreover, these norms are most successful when individuals coordinate on a shared standard for evaluating others' actions (Hilbe et al. 2018).

The experiment presented in this paper introduces two key innovations to the experimental literature on indirect reciprocity. First, rather than inferring norms solely from observed behavior (i.e., descriptive norms), I explicitly elicit both personal and social injunctive norms. Participants report what they personally believe is appropriate in a given situation, as well as what they believe most others consider appropriate. This dual elicitation offers deeper insights into the normative expectations that inform reputational judgments. Second, the study incorporates a coordination stage in which participants indicate which behaviors they believe should be considered appropriate for the group in specific contexts. The resulting group-level assessments—based on majority views—are then either shared with all participants or withheld, depending on the treatment condition. While the availability of this group norm information is exogenously determined by the experimenter, the content is endogenously generated, as it reflects participants' own collective input rather than externally imposed rules. This mechanism allows for the endogenous emergence of shared norms. The experiment employs a 2×2 design, varying both the cost of cooperation (*Low* vs. *High*) and the availability of group norm information (*NoInfo* vs. *Info*). The cost manipulation builds on prior findings that descriptive norms are sensitive to the costliness of cooperation. This study extends that inquiry by investigating whether cooperation costs also affect injunctive beliefs—both personal and perceived social. This design enables a detailed examination of how normative beliefs are shaped by the difficulty of the cooperation problem, and whether access to group norm information fosters greater behavioral alignment and supports higher levels of cooperation. The central research questions are therefore: (1) *Are personal, perceived social, and group injunctive norms sensitive to the cost of cooperation?* and (2) *Does the availability of group norm information promote convergence in behavior?*

Returning to the introductory example, previous studies observed how individuals like Alice and Dave responded to the information of others, but did not ask what they personally believed was the appropriate course of action, nor what they thought others considered appropriate. Moreover, Alice and Dave had no opportunity to express or coordinate on shared standards of evaluation. As a result, it remains unclear whether human subjects are able to align on a common reputational norm when given the opportunity to articulate their views, and coordinate with others. This study addresses that gap by explicitly eliciting normative beliefs and enabling group-level norm formation. In doing so, it offers a richer understanding of how indirect reciprocity is shaped by—and potentially reinforced through—shared normative expectations.

The remainder of this paper is structured as follows. Section 2 reviews the relevant theoretical and experimental literature on indirect reciprocity. Section 3 describes the experimental design. Section 4 introduces the norm framework and outlines the hypotheses. Section 5 presents the results. Finally, Section 6 provides a summary and concluding discussion.

2 Related literature

Theoretical studies

Assessment rules—the way actions are turned into reputations—have been central to the literature on indirect reciprocity since the introduction of “Image Scoring” by Nowak and Sigmund (1998). This rule is notable for its simplicity: it evaluates a donor’s reputation solely based on their observed actions, without considering the reputation of the recipient. As such, Image Scoring is classified as a first-order assessment rule. While Nowak and Sigmund (1998) demonstrated that Image Scoring can sustain cooperation under indirect reciprocity, Leimar and Hammerstein (2001) criticized the rule for failing to incentivize the withholding of cooperation from individuals with a bad reputation, as doing so can negatively impact the donor’s own reputation. They also showed that cooperation under Image Scoring is evolutionarily stable only under restrictive conditions. As an alternative, they proposed a rule similar to the “Standing” rule by Sugden (1986)—a second-order assessment rule in which reputation updates depend not only on the donor’s actions but also on the recipient’s reputation. Ohtsuki and Iwasa (2004) identified Standing as part of the leading eight, a set of assessment rules supporting high and stable cooperation under indirect reciprocity. These rules share a key principle: individuals maintain a good reputation by cooperating with well-reputed partners, while not being penalized for withholding cooperation from those with a bad reputation. Image Scoring is not part of the leading eight, all of which rely on at least second-order information by considering both the donor’s actions and the recipient’s reputation. Among the leading eight, “Stern Judging” has attracted particular attention (e.g., Pacheco et al. 2006; Santos et al. 2018). Unlike most other rules in this set, it requires refusing cooperation to those holding a bad reputation in order to maintain good standing. It has been shown that Stern Judging is at least as effective at sustaining high levels of cooperation compared to several more complex leading eight rules that rely on higher orders of information, such as third-order information—that is, rules that consider not only the donor's action and the recipient's reputation but also the donor's own reputation at the time of the action.

Experimental studies

Experimental studies initially explored indirect reciprocity by adapting the simulation framework of Nowak and Sigmund (1998) to investigate behavior of human subjects. For example, Wedekind and Milinski (2000) implemented an experimental version of the Helping Game, the stage game analyzed by Nowak & Sigmund. In this two-player game, already described above, a donor could choose to give a benefit to a passive receiver at a personal cost smaller than the benefit, or refuse to give (keep), which incurred no costs or benefits. Reputation was introduced by providing donors with the receiver's history of giving as a donor, based solely on first-order information, thus excluding any details about the context of these decisions. Seinen and Schram (2006) and Engelmann and Fischbacher (2009) employed similar designs. All of these studies found that indirect reciprocity is able to sustain cooperation among human subjects. In Seinen and Schram (2006), e.g., this is indicated by higher cooperation rates in treatments where reputation information is provided compared to those where it is not. However, because only first-order information was available, these experiments offer no insights into preferences for norms other than Image Scoring.

Milinski et al. (2001) addressed this by introducing second-order information. This means that donors not only saw their receiver's history of giving as a donor but also the giving history of the receiver's co-players for each donation decision. This expanded information allowed donors to evaluate whether refusing to give to a receiver was justified, enabling the adoption of the Standing rule. In particular, the study compared observed behavior with predictions under the Image Scoring and the Standing rule. The results showed that justified defection was rewarded much less than expected, with behavior aligning more closely with Image Scoring. The authors suggest that a possible limitation of the design was that the comprehensive information provided may have overwhelmed participants, causing them to rely mainly on first-order information instead.

Bolton et al. (2005) employed a similar design, but they limited information to the most recent relevant period. Under first-order information, donors saw only the receiver's last action as a donor. With second-order information, donors observed the receiver's last action as a donor and also the most recent action of their receiver in that interaction, giving rise to four possible distinct signals (e.g., *Give* to someone who chose *Keep* before). Relatively high giving rates following histories of refusal to give are interpreted as evidence that subjects do not necessarily follow the Image Scoring norm and take the context of a decision into account. However, these rates dropped sharply when the cost of giving increased. Additionally, average giving rates after refusals to give did not differ significantly between first- and second-order information treatments.

Swakman et al. (2016) investigated whether participants demand more context information when making decisions. In their setup, donors initially received first-order information about the last three decisions of the receiver. Donors could then choose to acquire second-order information, either for free or at a cost depending on the treatment. Many donors opted for this additional information, especially when the receiver's history indicated a refusal to give, suggesting the application of norms other than Image Scoring. Indeed, when second-order information was freely available, many donors used it to adopt strategies like Standing that account for justified defections. However, this behavior significantly decreased when acquiring information incurred a cost, and, overall, the most commonly applied norm was Image Scoring.

Gaudeul et al. (2021) used a design similar to the second-order information treatment in Bolton et al. (2005), but required subjects to make decisions using the strategy method in some rounds. Donors indicated whether to play *Give* or *Keep* for each of the four potential histories of receiver behavior, allowing for a more detailed analysis of underlying norms. The most commonly used strategy was to give to a receiver who had given before, aligning with Image Scoring. This was followed by unconditional giving, which reflects the incentives for maintaining a good reputation when the expectation is that others follow Image Scoring. Both strategies remained relatively stable regardless of the costs of giving, together accounting for about 60% of reported strategies. When giving costs were low, about 12% of participants adopted a strategy consistent with Standing, only withholding help when a receiver had defected with someone who had given before. However, this share dropped by half when costs increased, while the share of unconditional defectors rose sharply (to about 15%). Notably, very few participants chose a strategy consistent with Stern Judging, regardless of the costs.

Yamamoto et al. (2020) adopted an approach focused on the injunctive nature of norms. In their study, subjects were presented with various hypothetical scenarios in a helping game—e.g., a donor giving to a receiver with a good reputation or refusing to give to a receiver with a bad reputation—and then asked to evaluate the donor's behavior. The results indicate that subjects generally evaluated donors who gave very positively, regardless of the receiver's reputation. Evaluations of refusals to give were more nuanced: subjects strongly disapproved of donors who refused to give to someone with a good reputation but delivered no clear evaluation when the refusal was directed at someone with a bad reputation. This pattern does not align with any established assessment rules. It is important to note that subjects did not participate in the game themselves, and their evaluations were not incentivized.

3 Experimental Design

This study seeks to address two gaps in the literature on indirect reciprocity. First, it extends prior work by eliciting not only descriptive but also personal and perceived injunctive norms. Second, it introduces a coordination stage to examine how group norm communication influences norm adoption and cooperation. The findings offer new insights into the normative foundations of indirect reciprocity and the role of coordination in supporting more cooperative norms. Following the literature, I study indirect reciprocity within the framework of the Helping Game. The experimental design manipulated two factors: the cost of giving and the information provided about group norms. The gameplay was structured into two phases, each consisting of six rounds of the Helping Game. Between the two phases, participants' preferences regarding personal injunctive norms, beliefs about social injunctive norms, and preferences for group norms were elicited. Depending on the treatment, this information was made public before Phase 2 or not. An illustration of the game phases is shown in Figure 1

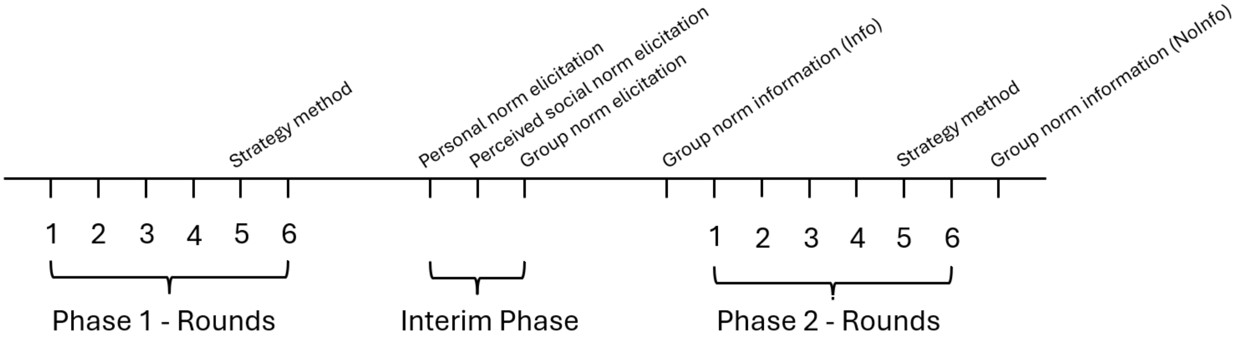

Strategy method
Personal norm elicitation
Perceived social norm elicitation
Group norm elicitation
Group norm information (Info)
Strategy method
Group norm information (NoInfo)
1
2
3
4
5
6
1
2
3
4
5
6
Phase 1 - Rounds
Interim Phase
Phase 2 - Rounds

Figure 1: **Overview of game phases**

Notes: Players play six rounds of the Helping Game in Phase 1; decision-making is conducted via the strategy method in Round 5, otherwise via direct elicitation. In the Interim Phase, different types of norms are elicited. Depending on the treatment, the group norm is made public before or after Phase 2. Gameplay in Phase 2 is otherwise the same as in Phase 1.

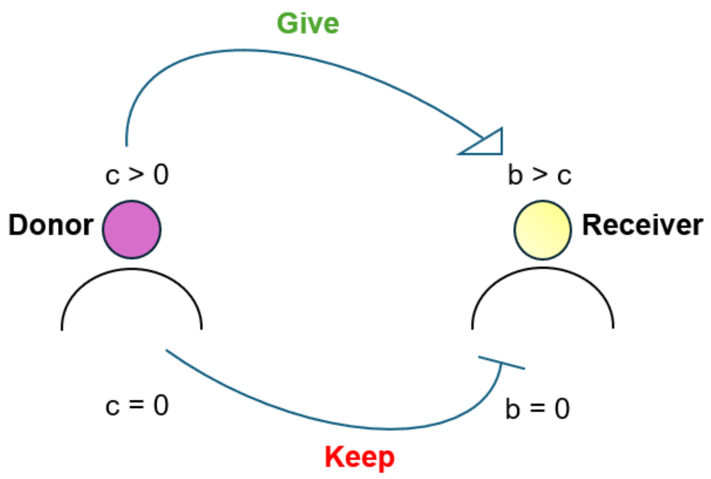


Figure 2: **Overview of the Helping Game**

Notes: The donor decides between *Give* and *Keep*. If the donor chooses *Give*, she generates a benefit b for the receiver at a cost c to herself. If she chooses *Keep*, no cost or benefits are generated.

Helping game

The Helping Game was played in pairs. In each pair, one participant was randomly assigned the role of donor, and the other was assigned the role of receiver.[2] The donor had to choose between two actions: *Give* and *Keep*.[3] The receiver had no decision to make. If the donor chose *Give*, they incurred a cost, c, to generate a benefit, b, for the receiver, where $b > c > 0$. If the donor chose *Keep*, no costs or benefits were incurred. Given the ratio of costs to benefits, choosing *Give* increases social welfare at a personal cost to the donor. Figure 2 illustrates the key aspects of the game.

[2] In the experiment, the "donor" was referred to as "decider;" in this paper, I stick to donor as the term is generally used in the literature.

[3] In line with the literature, e.g., Engelmann and Fischbacher (2009), the choices were not neutrally framed and were presented as *Give* and *Keep* to the subjects.

Matching and roles

Participants engaged in a series of one-shot interactions of the Helping Game, with no two players matched more than once (i.e., perfect stranger matching).[4] While each interaction involved only one donor and one receiver, all players were required to make a decision as if they were the donor in each interaction. After both players had made their choices, the decision of one randomly selected player was implemented, while the other player's choice had no impact on the payoffs. Participants were informed of the outcome of the random draw, their co-player's decision, and their own payoff only after the experiment had concluded.

Reputation information

To facilitate indirect reciprocity as a mechanism of cooperation, players carried over reputation information from one interaction to the next. In each round, both players in a pairing received reputation information about their co-player before making their decision. The amount of information provided increased progressively up to the third round of play. In the first round, players received no information about their co-player. In the second round, players were informed whether their co-player chose *Give* or *Keep* in the first round. In the third round, players received information about their co-player's decision in the second round, as well as the decision of the second-round interaction partner of their co-player in the first round. In the fourth round, players were informed about their co-player's decision in the third round, along with the decision of the third-round partner of their co-player in the second round.[5] This pattern continued in subsequent rounds. Beginning in the third round, each player thus received one of four possible signals about their current co-player:

[4] Perfect stranger matching appears most appropriate for capturing the spirit of indirect reciprocity and is used in comparable studies (e.g., Gaudeul et al. 2021). However, there are also studies on indirect reciprocity that use different matching protocols. For example, Swakman et al. (2016) employed random re-matching, which enables a greater number of interactions within a single session. However, such designs may introduce elements of direct reciprocity, as participants can interact with the same partners more than once, potentially confounding the indirect reciprocity mechanism.

[5] In all cases, the reputation information was based on the choices, regardless of whether the choice had actually been implemented and was payoff-relevant or not.

- **S1 - Give/Give:** The current co-player chose *Give* in the previous round when confronted with someone who chose *Give* before.

- **S2 - Keep/Give:** The current co-player chose *Keep* in the previous round when confronted with someone who chose *Give* before.

- **S3 - Give/Keep:** The current co-player chose *Give* in the previous round when confronted with someone who chose *Keep* before.

- **S4 - Keep/Keep:** The current co-player chose *Keep* in the previous round when confronted with someone who chose *Keep* before.

Figure A1 in the Appendix illustrates how reputation information was presented to the participants. The provided information can be described as incomplete second-order information. While it offered some context for a player's decision, it only included details from up to two previous rounds and did not enable players to unambiguously determine whether a decision was justified. For instance, if a player received the signal "S4: Keep/Keep" in Round 4, it might have indicated a justified defection by the current co-player (choosing *Keep* in Round 3 because their previous co-player chose *Keep* in Round 2). However, it remained unclear whether the decision by the earlier co-player might have been justified itself, because their former co-player played *Keep* in Round 1 (see Section 4 for more details on this). The design choice to use this particular kind of reputation information, despite its ambiguities, was based on three main considerations: First, this kind of reputation information is used in related studies (e.g., Bolton et al. 2005; Gaudeul et al. 2021), which facilitates comparability of findings. Second, limiting the number of signals reduces cognitive load and avoids overwhelming participants. Third, these signals mirror real-life situations, where people often make moral judgments with limited information about others' past behavior.

Phase 1

Phase 1 of the experiment consisted of six rounds. Prior to this phase, participants received only the minimal information necessary for making informed decisions, without any additional explanations or norm elicitation tasks that could have primed behavior. Following Gaudeul et al. (2021), the strategy method was used for decision-making in a selected round (here: Round 5). In this round, players had to make one choice for each of the four possible co-player reputation signals, and their decision for the actual situation was then implemented.[6] This approach allowed to observe players' preferred behavior in each potential situation, thereby inferring their personal descriptive norm. Decisions for Rounds 1, 2, 3, 4, and 6 were elicited using the direct response method. The instructions, quiz, and the main decision screen are included in Section 9 of the Appendix.

Interim phase

After Round 6 of Phase 1, participants entered the Interim Phase, during which their personal injunctive norms, perceived social injunctive norms, and group norms were elicited.

Personal injunctive norms

For the elicitation of personal injunctive norms, participants were presented with each of the four potential reputation signals described above. For each signal, they were asked to indicate which behavior (*Give* or *Keep*) they personally consider appropriate in the sense of being "right" or "moral," independent of others' opinions. The wording closely followed Bašić and Verrina (2024), who study personal and social norms in the context of social preferences. To emphasize the injunctive nature of the norm, participants were explicitly reminded to respond independently of their actual or intended behavior in the game. Section 10 of the Appendix provides an example for the first signal.

[6] Before making decisions via the strategy method, participants received a detailed explanation including examples and completed a quiz to ensure their understanding.

Perceived social injunctive norms

For the elicitation of perceived social injunctive norms, participants responded to a simple “Yes” or “No” question indicating whether they believed their personal injunctive norm aligned with the majority opinion within their session. This question was asked immediately after participants indicated their personal injunctive norm for a given signal. For instance, if a participant deemed *Keep* appropriate following “S3: Give/Keep” but believed that the majority considered *Give* appropriate, they should have responded “No.” The combination of the participant’s personal norm and their belief about the majority’s norm was then used to construct their perceived social injunctive norm. To incentivize accurate responses, participants earned 1.20€ for each correct answer; in the example above, the participant would have received this reward if the majority in the session indeed judged *Give* as appropriate for that signal. This method for eliciting incentivized second-order beliefs is related to the approach established by Krupka and Weber (2013). An example related to the first signal is provided in Section 10 of the Appendix.

Group norms

In the final stage of the Interim Phase, group norms were elicited. Participants were introduced to group norms as the behavior they would recommend to the entire group in response to a specific signal, with the goal of fostering a shared understanding of appropriate actions in that context. This separate elicitation of group norms—rather than relying solely on personal and social injunctive norms—allows for a deeper insight into how participants perceive norms collectively and which behaviors they consider most successful in practice. For example, a participant may personally view a certain action as appropriate but believe it to be too complex or difficult for the group to consistently follow, or may think that an alternative behavior is more effective at promoting cooperation. To reduce cognitive load, group norms were elicited for only two of the four possible signals: “S3: Give/Keep” and “S4: Keep/Keep.” These are also the signals to which the three norms under focus—Image Scoring, Standing, and Stern Judging—assign different prescriptions. For both signals, participants were presented with one argument supporting the choice of *Give* and one supporting the choice of *Keep*, each framed according to the moral principles underlying Image Scoring, Standing, and Stern Judging. For example, for "S3: Give/Keep" the argument for playing *Give* was: "Person C should choose *Give*. This is appropriate because Person B has previously chosen *Give* themselves." For playing *Keep* the argument was: "Person C should choose *Keep*. Although Person B has previously chosen *Give*, it would have been appropriate to choose *Keep* instead because Person A had previously chosen *Keep*." Section 10 of the Appendix includes the exact example.

Phase 2

Gameplay in Phase 2 proceeded identically to Phase 1. No reputation information from Phase 1 was carried over, which provided participants with a fresh start. As in Phase 1, decisions were made using the strategy method in Round 5 and by the direct response method in all other rounds.

Treatments

The experiment investigated whether the difficulty of cooperation influences norm preferences and if a norm coordination device affects game behavior. To achieve this, I employed a 2 × 2 design, manipulating both the cost of giving and the availability of a group norm coordination device. Table 1 provides an overview of the treatments.

Table 1: **Overview of treatments**

	Group Norm Info = No	**Group Norm Info = Yes**
Cost of Giving = 15	Low_NoInfo	Low_Info
Cost of Giving = 45	High_NoInfo	High_Info

Treatment variation: cost of giving

In all treatments, when a donor selected *Keep*, they received 100 LabDollars (LD), while the receiver earned 50 LD. In the low-cost treatments, the donor received 85 LD when choosing *Give*, while the receiver earned 125 LD. In the high-cost treatments, the donor received 55 LD when choosing *Give*, while the receiver earned 125 LD. This means the cost of giving was either $c_{Low} = 15$ or $c_{High} = 45$, while the benefit of giving remained constant at $b = 75$. Consequently, the cost-to-benefit ratios were $c/b_{Low} = 1/5$ for the low-cost treatments and $c/b_{High} = 3/5$ for the high-cost treatments. The ratios were chosen to match those used in the corresponding treatments in Bolton et al. (2005). The variation in the cost of giving applied from Phase 1 onward for all rounds of play and was also relevant for the Interim Phase.

Treatment variation: group norm information

This treatment variation manipulated whether participants received the outcome of the group norm elicitation before the start of Phase 2 or after. In the *Info* treatments, participants were informed about the majority-recommended behavior (*Give* or *Keep*) for the signals "S3: Give/Keep" and "S4: Keep/Keep" immediately before Phase 2 began. Specifically, participants saw the relevant signal again (e.g., "S3: Give/Keep") along with the text: "The majority of participants indicate *Give* as the non-binding behavioral recommendation for this situation," if *Give* was the majority choice. In contrast, participants in the *NoInfo* treatments received this information only after completing Phase 2, preventing its use as a coordination device during the game. At the time of group norm elicitation, participants in both treatments were informed only that they would receive the group norm outcome either before or after Phase 2, without exact timing, ensuring comparable conditions during elicitation. Section 10 of the Appendix includes an example of how the information was presented.

Payment

The total payment for a participant consisted of earnings from each round of play in Phases 1 and 2, and four payments for the perceived social injunctive norm elicitation in the Interim Phase. Earnings in Phases 1 and 2 were determined by the participant's own choice or their co-player's choice, based on the outcome of the random role assignment. Payment in the Interim Phase depended on the accuracy of their beliefs about perceived social injunctive norms. The conversion rate from LD to € was 1 LD = 0.015€ (1.5 euro cents).

Procedures

Between December 2024 and July 2025, a total of 336 participants took part in 24 experimental sessions conducted in the computer lab KLab at the University of Kassel, Germany. Each session implemented one of the four treatments, with every participant assigned to only one treatment, following a between-subjects design. Undergraduate students from various academic backgrounds were recruited through ORSEE (Greiner 2015). The experiment was programmed in oTree (Chen et al. 2016). At the beginning of each session, participants were randomly assigned to cubicles, and printed instructions were distributed. The experimenter then read the instructions aloud to ensure clear understanding of the rules. The session proceeded with a quiz to check comprehension. Phase 1 commenced only after all participants had answered the quiz correctly. Following Phase 2, participants completed a questionnaire about their personal characteristics and their behavior during the game. Sample characteristics and randomization tests are reported in Table A1 of the Appendix, these results indicate that there are no significant differences between treatments with respect to the elicited characteristics. After completing the questionnaire (provided in Section 11 of the Appendix), participants received their total earnings in cash. Sessions lasted between 70 and 90 minutes. Earnings ranged from approximately 14€ to 25€, with an average of about 20€. No show-up fee was provided. The study received ethical approval from the German Association for Experimental Economic Research (GfeW). The approval documentation is available at: https://gfew.de/ethik/M4vRQ7jC.

4 Norms and Research Questions

Norms

Besides giving behavior, elicited norms are the main outcome variables of this study. Throughout this paper, a norm is defined as the pattern of responses a subject provided when presented with different signals. In this section, I first explain how these norms were derived and classified. I then clarify how this definition differs from the concept of norms in the theoretical literature—referred to here as assessment rules. This explanation draws on the framework and discussion provided in Gaudeul et al. (2021), who were the first to use the strategy method to elicit norms in the context of experimental studies on indirect reciprocity.

As outlined, at different stages of the experiment, subjects encountered all or some of the four potential signals and were asked to indicate: (1) what action they wanted to take (personal descriptive norm), (2) what action they personally deemed appropriate (personal injunctive norm), (3) what action they believed the majority of other players considered appropriate (perceived social injunctive norm), and (4) what action they wished to recommend to the entire group (group norm). For personal and perceived social injunctive norms, subjects responded to all four signals, choosing either *Give* or *Keep* for each, which potentially yielded 16 distinct norm profiles. For the group norm, only two signals were evaluated, with participants indicating their recommended behavior for each, resulting in four possible norms. Based on their responses, the player's behavior was classified using the norms defined below. For example, suppose that in Round 5 of Phase 1 (elicitation via the strategy method), a player indicated behavior in line with the Image Scoring rule, i.e., the following choices in response to the four signals:

Play *Give* (= 1) when the co-player played *Give* in the previous round and play *Keep* (= 0) otherwise. The Phase 1 personal descriptive norm (PD) of this player can be represented as:

$$PD_1 = (1, 0, 1, 0),$$

where the first digit indicates *Give* after "S1: Give/Give," the second digit indicates *Keep* after "S2: Keep/Give," the third digit indicates *Give* after "S3: Give/Keep," and the fourth digit indicates *Keep* after "S4: Keep/Keep." Table 2 provides an overview of all 16 potential norms that can emerge when four signals are presented to subjects, while Table 3 displays the potential norms based on two signals.

The assessment rule column in Table 2 & Table 3 indicates the name that is used in the theoretical literature for the assessment rule most closely related to the given norm. In the theoretical literature on indirect reciprocity, an assessment rule determines how a donor's binary reputation (GOOD or BAD) is updated (e.g., Ohtsuki and Iwasa 2004). This can be based solely on the donor's action (first-order rule), on the donor's action and the receiver's reputation (second-order rule), or on the donor's action along with both the receiver's and donor's reputations (third-order rule) or even higher orders of information.[7] For example, Image Scoring is a first-order assessment rule: a donor's reputation will be GOOD if they play *Give* and BAD if they play *Keep*, regardless of the receiver's reputation. Simple Standing is a second-order assessment rule, where a donor receives a GOOD reputation if they play *Give* (regardless of the receiver's reputation) or if they play *Keep* when interacting with a receiver who has a BAD reputation. Stern Judging is also a second-order assessment rule, assigning a donor a GOOD reputation when they play *Give* with someone who is GOOD or *Keep* with someone who is BAD, and a BAD reputation otherwise. Assessment rules are combined with action rules to form a strategy. An action rule determines whether an

[7] In this study, I only consider assessment rules up to the second order. Therefore, e.g., I refer to the Simple Standing rule and not other Standing rules where the evaluation of a defection against a receiver in bad reputation depends on the reputation of the donor.

individual plays *Give* or *Keep* based on the recipient's reputation. A common action rule is to play *Give* when interacting with someone who is GOOD and *Keep* when interacting with someone who is BAD.

It is important to note that what is referred to as a norm in this paper is not directly comparable to the assessment rules discussed in the theoretical literature. The primary difference is that subjects did not directly carry or observe reputations that were automatically updated according to a fixed assessment rule; instead, they observed reputations only indirectly through signals and updated reputations according to their own judgment. Moreover, these signals were incomplete and did not allow for the unambiguous inference of a reputation except when applying a first-order norm. This can be illustrated using the example in Figure 3:

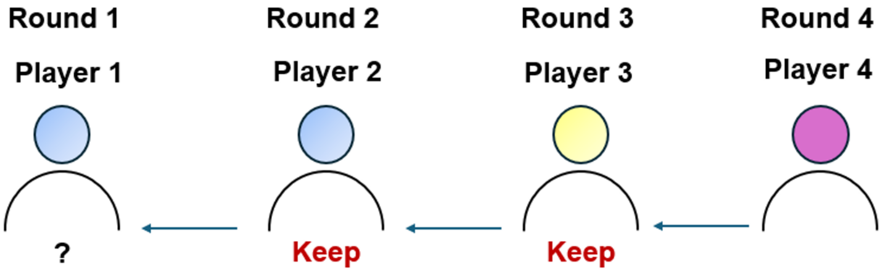


Figure 3: **Example for "S4: Keep/Keep"**

Notes: Under the incomplete information used in this experiment, the reputation of P3 cannot unambiguously be assessed by P4 as the signal "S4: Keep/Keep" in this context leads to GOOD reputation when the choice of P1 was *Give* and BAD reputation when the choice of P1 was *Keep*.

In Round 4, Player 4 receives the signal "S4: Keep/Keep" for their current co-player, Player 3. Under the Simple Standing assessment rule, Player 3's reputation depends on their own action as well as the actions of Players 2 and 1. For Player 3's *Keep* after *Keep* to result in a GOOD reputation, Player 1 must have chosen *Give*. In this specific case, Player 2's *Keep* would have led to a BAD reputation for Player 2. Since responding with *Keep* to a BAD reputation results in a GOOD reputation under Simple Standing, Player 3 would then be considered GOOD. Conversely, if Player 1 had chosen *Keep*, Player 2 would have retained a GOOD reputation. In this scenario, Player 3's *Keep* against a co-player with a GOOD reputation results in a BAD reputation, marking Player 3 as BAD.

While the above example illustrates that the norms in this experiment are not directly comparable to assessment rules, specific norms can be argued to carry their moral content. For example, the norm $(1, 0, 1, 1)$ prescribes *Give* after "S1: Give/Give," "S3: Give/Keep," "S4: Keep/Keep," and *Keep* otherwise. This is comparable to the Simple Standing assessment rule, combined with the action rule of playing *Give* with a GOOD reputation and Keep otherwise. Similarly, the norm $(1, 0, 0, 1)$ prescribes *Give* after "S1: Give/Give" and "S4: Keep/Keep," while prescribing *Keep* after "S2: Keep/Give" and "S3: Give/Keep." This is related to the Stern Judging assessment rule, combined with the action rule of playing *Give* with a GOOD reputation and *Keep* otherwise.

Table 2: **Norms and associated assessment rules based on four signals.**

Signal & Choice					
S1: Give/Give	**S2: Keep/Give**	**S3: Give/Keep**	**S4: Keep/Keep**	**Norm**	**Assessment Rule**
Give	Give	Give	Give	(1, 1, 1, 1)	ALLC
Keep	Keep	Keep	Keep	(0, 0, 0, 0)	ALLD
Give	Keep	Give	Keep	(1, 0, 1, 0)	Image Scoring
Give	Keep	Give	Give	(1, 0, 1, 1)	Simple Standing
Give	Keep	Keep	Give	(1, 0, 0, 1)	Stern Judging
Give	Keep	Keep	Keep	(1, 0, 0, 0)	Shunning
Give	Give	Give	Keep	(1, 1, 1, 0)	Mild Shunning
Give	Give	Keep	Give	(1, 1, 0, 1)	
Keep	Give	Give	Give	(0, 1, 1, 1)	
Give	Give	Keep	Keep	(1, 1, 0, 0)	
Keep	Give	Give	Keep	(0, 1, 1, 0)	
Keep	Give	Keep	Give	(0, 1, 0, 1)	
Keep	Keep	Give	Give	(0, 0, 1, 1)	
Keep	Give	Keep	Keep	(0, 1, 0, 0)	
Keep	Keep	Give	Keep	(0, 0, 1, 0)	
Keep	Keep	Keep	Give	(0, 0, 0, 1)	

Notes: The table shows all possible combinations that can be derived from four signals, with two possible actions (*Give* or *Keep*) in response to each. If a player indicates, e.g., *Give* after signals S1, S3, and S4, and *Keep* otherwise, their norm is represented as (1, 0, 1, 1). The corresponding assessment rule refers to the classification associated with this norm in the theoretical literature. The term "Mild Shunning" was first used by Gaudeul et al. (2021) to describe this specific norm.

Table 3: **Norms and associated assessment rules based on two signals.**

Signal & Choice			
S3: Give/Keep	**S4: Keep/Keep**	**Norm**	**Assessment Rule**
Give	Give	(1, 1)	Simple Standing
Keep	Keep	(0, 0)	ALLD
Give	Keep	(1, 0)	Image Scoring
Keep	Give	(0, 1)	Stern Judging

Notes: The table shows all possible combinations that can be derived from two signals, with two possible actions (*Give* or *Keep*) in response to each.

Hypotheses

The primary focus of this study is to examine the types of norms elicited during the Interim Phase and their relationship to the cost of giving. In particular, the analysis explores whether participants tended to favor norms that reward justified defections—specifically, norm profiles such as $(1, 0, 1, 1)$ (associated with Simple Standing) or $(1, 0, 0, 1)$ (associated with Stern Judging)—over norms that do not, such as $(1, 0, 1, 0)$ (associated with Image Scoring). The specific hypotheses are listed and described below. All hypotheses, along with the sample size and key design elements, were preregistered prior to any data collection. The full preregistration is available at: https://aspredicted.org/cct6-crch.pdf.

Hypothesis 1: If the cost of giving is high, participants are more likely to indicate a personal injunctive norm that rewards justified defections (i.e., playing *Keep* after *Keep*), resulting in a higher prevalence of norms related to Simple Standing and Stern Judging and a lower prevalence of the norm related to Image Scoring.

Hypothesis 2: If the cost of giving is high, participants are more likely to indicate a perceived social injunctive norm that rewards justified defections (i.e., playing *Keep* after *Keep*), resulting in a higher prevalence of norms related to Simple Standing and Stern Judging and a lower prevalence of the norm related to Image Scoring.

Hypothesis 3: If the cost of giving is high, participants are more likely to indicate a group norm that rewards justified defections (i.e., playing *Keep* after *Keep*), resulting in a higher prevalence of norms related to Simple Standing and Stern Judging and a lower prevalence of the norm related to Image Scoring.

The rationale for Hypotheses 1 through 3 is as follows: higher costs of giving are expected to increase the rate of non-cooperative behavior. As defection becomes more common, the need for effective and justifiable punishment mechanisms also increases. Norms such as Simple Standing and Stern Judging, which allow for justified defections by using *Keep* directed at players who chose *Keep* before, may be perceived as more effective and thus more appropriate under high-cost conditions. In contrast, Image Scoring penalizes all defections indiscriminately, which could make it less appealing in such contexts.

Hypothesis 4 refers to the use of the coordination device introduced in Phase 2, which allows participants to observe the group norm elicited during the Interim Phase. It examines whether making this norm public influences subsequent behavior by serving as a coordination mechanism.

Hypothesis 4: Participants are more likely to behave in accordance with the group norm when that norm is made public before Phase 2.

5 Results

In this section, I present the results in terms of giving rates, norm preferences, and norm alignment. The results are organized according to the different phases of the experiment. The information variation became relevant in Phase 2 only, and I do not find significant differences in Phase 1 giving rates between *Low_NoInfo* and *Low_Info* ($p = 0.933$) or between *High_NoInfo* and *High_Info* ($p = 0.814$).[8] Accordingly, for all results prior to Phase 2, I pool low- and high-cost conditions and do not differentiate between *NoInfo* and *Info* treatments.

5.1 Phase 1 - Giving rates and personal descriptive norms

Panel (a) of Figure **??** shows the average proportion of *Give* choices per round in Phase 1. Across all rounds—including both direct decisions and implemented choices elicited via the strategy method—participants chose to play *Give* in 65.48% of interactions when costs were low. This indicates that indirect reciprocity is able to promote a meaningful level of cooperation. However, cooperation is far from universal even under relatively low costs. Furthermore, higher cooperation costs resulted in a significant drop in *Give* choices, which decreased to 54.37% ($p = 0.001$ compared to *Low*).

Panel (b) of Figure **??** shows decisions conditional on the different signals for Rounds 3, 4, and 6 of Phase 1. The results indicate that participants in both cost conditions were significantly more likely to give when their receiver previously gave and less likely to give when the receiver previously chose to keep ($p < 0.001$ for *Low*; $p < 0.001$ for *High*). Subjects appear not to take the decision context into account: within each cost condition, there is no significant difference in the share of *Give* choices between signals "S1: Give/Give" and "S3: Give/Keep" ($p = 0.228$ for *Low*; $p = 0.600$ for *High*), or between "S2: Keep/Give" and "S4: Keep/Keep" ($p = 0.247$ for *Low*; $p = 0.709$ for *High*).[9]

[8] Based on binary probit models with the decision ($Give = 1$, $Keep = 0$) as the dependent variable and an indicator for the treatment as the explanatory variable; standard errors clustered at the individual level. Random-effects panel probit models yield qualitatively similar findings.

[9] All p-values in this paragraph are based on binary probit models with the decision ($Give = 1$, $Keep = 0$) as the dependent variable and indicators for either the cost condition, the previous action of the co-player (*Give* or *Keep*), or the signal (e.g., S1 or S3) as explanatory variable; standard errors clustered at the individual level. Random-effects panel probit models yield qualitatively similar findings

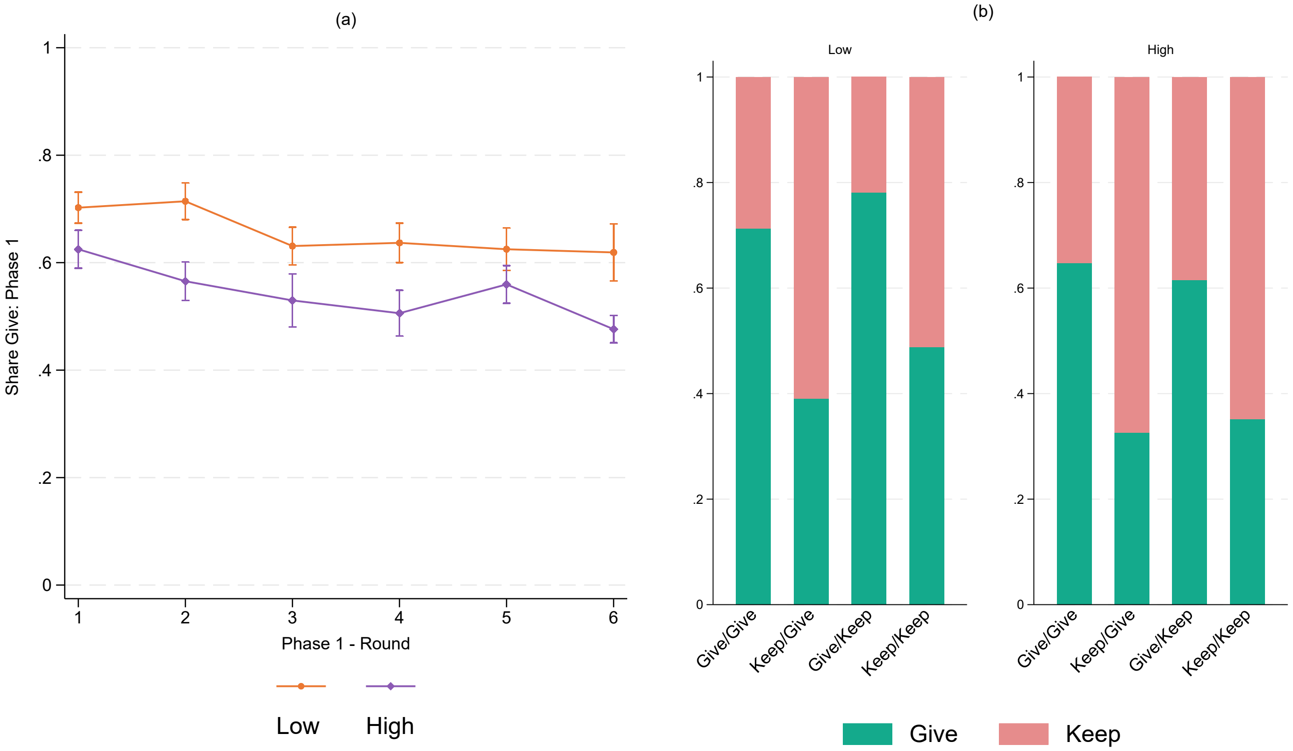


Figure 4: **Giving rates by treatment and signal (Phase 1)**

Notes: *Panel (a)* Markers indicate average share of *Give* by round of Phase 1. For Round 5 (strategy method), the implemented choice was used. Orange (circle) markers indicate the pooled low-cost treatments (*Info* + *NoInfo*). Violet (diamond) markers indicate the pooled high-cost treatments (*Info* + *NoInfo*). Whiskers indicate standard errors clustered at the session level. *Panel (b)* Bars indicate shares of *Give* and *Keep* by signal for Round 3, 4 and 6 of Phase 1 for low-cost treatments (left) and high-cost treatments (right).

While these results suggest that participants seem to evaluate reputations based on actions and do not use norms that explicitly distinguish between justified and unjustified behavior, more consistent and detailed insights into their personal descriptive norms emerge from the strategy method responses in Round 5.[10] Unlike in other rounds, where participants only encountered a subset of possible signals depending on their matchings, the strategy method ensured that each subject made a decision for every possible signal. Table 4 reports the

[10] While the analysis aggregates individual responses and thus reflects social norms, I use the term personal norms to emphasize that these are derived from individual evaluations, as elicited through the strategy method.

share of *Give* choices for each signal (left section) and the corresponding norms derived from individual choices (right section). The outcomes qualitatively align with those from the direct elicitation: in both cost conditions, subjects were more likely to choose *Give* when the co-player previously gave, and more likely to choose *Keep* otherwise. Under low costs, the giving rates following *Keep* signals are somewhat higher compared to the direct elicitation. By contrast, under high costs, the *Give* rates are very similar across the strategy method and direct elicitation. Turning to the derived norms, the results indicate that, in both cost conditions, the majority of participants (over 70%) selected norms consistent with assessment rules commonly discussed in the theoretical literature. However, a comparison across the low- and high-cost conditions reveals notable differences, particularly regarding (nearly) unconditional behavior. Participants in the low-cost conditions were significantly more likely to endorse norms that prescribe (nearly) unconditional giving—such as (1, 1, 1, 1), related to ALLC, and (1, 1, 1, 0), related to Mild Shunning. By contrast, participants in the high-cost conditions more frequently selected norms that prescribe (nearly) unconditional keeping—such as (0, 0, 0, 0), related to ALLD, and (1, 0, 0, 0), related to Shunning. Under high costs, about 14% of participants adopted behavior consistent with (1, 0, 1, 1), related to Simple Standing, which explicitly distinguishes between justified and unjustified defections; this share is significantly lower under low costs, maybe reflecting greater acceptance of retaliatory behavior when cooperation is more costly. There is no difference between cost conditions in the adoption of the norm (1, 0, 1, 0), related to Image Scoring. With an adoption rate of about 16% in both cost conditions, its overall prevalence is notably lower than the levels observed by Gaudeul et al. (2021). In contrast, the findings for the norm (1, 0, 0, 1), related to Stern Judging, are consistent across studies. With very low adoption rates, the results suggest that this norm plays no meaningful role in guiding behavior in this context.

Table 4: **Personal descriptive norms (Phase 1)**

(1) Giving rates			**(2) Norms**				
Signal	**Low**	**High**	**Norm**	**Assessment Rule**	**Low**	**High**	**p-value**
Give/Give	0.7202	0.6548	(1, 1, 1, 1)	ALLC	0.2738	0.1667	0.015
Keep/Give	0.5714	0.3571	(0, 0, 0, 0)	ALLD	0.0595	0.1250	0.014
Give/Keep	0.7679	0.6488	(1, 0, 1, 0)	Image Scoring	0.1607	0.1607	1.000
Keep/Keep	0.4821	0.4048	(1, 0, 1, 1)	Simple Standing	0.0774	0.1429	0.038
			(1, 0, 0, 1)	Stern Judging	0.0119	0.0060	0.551
			(1, 0, 0, 0)	Shunning	0.0238	0.1012	0.001
			(1, 1, 1, 0)	Mild Shunning	0.1250	0.0476	0.004
			Other		0.2679	0.2500	0.688

Notes: *(Left section)* Numbers indicate the share of respondents that selected *Give* as choice in response to the respective signal. *(Right section)* Numbers indicate the share of participants who selected each respective norm based on elicitation using the strategy method in Round 5 of Phase 1. Reported p-values are derived from binary probit models, where the dependent variable is an indicator for the chosen norm and the explanatory variable is an indicator for the cost condition with standard errors clustered at the session level. Number of observations = 168 per cost condition. The full set of norms is reported in Table A4 of the Appendix.

5.2 Interim phase - Personal injunctive norms

Descriptive norms reflect observed game behavior, but such behavior must not necessarily align with what individuals personally consider appropriate. To uncover these underlying normative judgments, it is essential to examine personal injunctive norms, which capture players' own beliefs about what one ought to do. Personal injunctive norms were elicited during the Interim Phase, after participants had gained experience with both the game and the strategy method. As outlined above, participants were asked to indicate, for each signal, the behavior they personally deemed appropriate—without reference to others' views and independent of actual behavior in the game. It is worth emphasizing that participants were presented with situations that could be assessed unambiguously (Round 3). The results of the personal injunctive norm elicitation are presented in Table 5.

Compared to the descriptive norms, the outcomes from the personal injunctive norms elicitation are markedly more consistent across cost conditions, with significant differences observed for only two of the norms under consideration. This suggests that injunctive judgments are less sensitive to cost variations than actual game behavior. Building on these findings, I now turn to Hypothesis 1, which examines whether participants exhibited a stronger preference for norms that reward justified defection under high-cost conditions. The results offer no support for this hypothesis: under high costs, 44.05% of subjects indicated (1, 0, 1, 0)—rooted in the Image Scoring rule—as most appropriate. This share is significantly lower under conditions of low costs at 29.17%. For (1, 0, 1, 1) and (1, 0, 0, 1), related to Simple Standing and Stern Judging respectively, on the other hand, there are no significant difference between cost conditions. Again, for the norm related to Stern Judging, the shares are extremely low in both cost conditions.

Table 5: **Personal injunctive norms**

(1) Giving appropriate			**(2) Norms**				
Signal	**Low**	**High**	**Norm**	**Assessment Rule**	**Low**	**High**	**p-value**
S1: Give/Give	0.9167	0.9226	(1, 1, 1, 1)	ALLC	0.2500	0.1369	0.008
S2: Keep/Give	0.4405	0.2798	(0, 0, 0, 0)	ALLD	0.0238	0.0298	0.718
S3: Give/Keep	0.8631	0.8810	(1, 0, 1, 0)	Image Scoring	0.2917	0.4405	0.001
S4: Keep/Keep	0.4405	0.3393	(1, 0, 1, 1)	Simple Standing	0.1369	0.1726	0.324
			(1, 0, 0, 1)	Stern Judging	0.0179	0.0119	0.675
			(1, 0, 0, 0)	Shunning	0.0476	0.0357	0.567
			(1, 1, 1, 0)	Mild Shunning	0.1369	0.1012	0.362
			Other		0.0952	0.0714	0.421

Notes: *(Left section)* Numbers indicate the share of respondents that selected *Give* as appropriate in response to the respective signal. *(Right section)* Numbers indicate the share of participants who selected each respective norm based on elicitation in the Interim Phase. Reported p-values are derived from binary probit models, where the dependent variable is an indicator for the chosen norm and the explanatory variable is an indicator for the cost condition with standard errors clustered at the session level. Number of observations = 168 per cost condition. The full set of norms is reported in Table A5 of the Appendix. Additional results from multinomial logit models that include control variables are presented in Table A6 of the Appendix.

Table A6 in the Appendix reports results from multinomial logit models that include additional control variables. Regarding the impact of the cost condition, the findings are consistent with those from the binary probit models presented here. The results further show that subjects who expressed stronger tendencies toward negative (direct) reciprocity—that is, a willingness to punish others for unkind or unfair behavior—were less likely to indicate (1, 1, 1, 1), related to ALLC, as appropriate. This suggests that preferences for direct punishment may also extend to indirect reciprocity contexts.

5.3 Interim Phase - Perceived social injunctive norms

Perceived social injunctive norms were elicited by asking participants whether they believed the majority of subjects in their session agreed with their own personal injunctive norm for each signal. Correct beliefs were incentivized with a payment of 1.20€ each. The results of this elicitation are presented in Table 6.

These findings again reveal strong alignment between cost conditions: there are no significant differences in perceived norm shares between the low- and high-cost treatments for any of the norms under investigation. This provides no support for Hypothesis 2, which posited that under high costs, participants would be more likely to perceive norms such as (1, 0, 1, 1), related to Simple Standing, and (1, 0, 0, 1), related to Stern Judging, as appropriate, and (1, 0, 1, 0), related to Image Scoring, as less appropriate. Instead, across both cost conditions, the majority of subjects believed that others considered behavior aligned with Image Scoring (1, 0, 1, 0) to be most appropriate. In contrast, perceived endorsement of Simple Standing (1, 0, 1, 1) and Stern Judging (1, 0, 0, 1) was very low in both conditions. When compared to the actual results from the personal injunctive norm elicitation in Table 6.5, these perceptions—particularly regarding (1, 0, 1, 0), rooted in Image Scoring—appear to be clear misconceptions. This is especially evident in the low-cost condition, where the actual share of participants personally endorsing the (1, 0, 1, 0) norm related to Image Scoring was much lower than perceived. Conversely, participants underestimated the prevalence of (1, 0, 1, 1), related to Simple Standing, as a personal norm. Furthermore, across both cost conditions, participants significantly underestimated the share of others who deemed unconditional giving (following the (1, 1, 1, 1) ALLC norm) to be appropriate. Table A8 in the Appendix reports results from multinomial logit models that include additional control variables. Regarding the impact of the cost condition, the findings are largely consistent with those from the binary probit models presented here; the only difference is a weakly significant effect of the cost condition for ALLC.

Table 6: **Perceived social injunctive norms**

(1) Giving appropriate			**(2) Norms**				
Signal	**Low**	**High**	**Norm**	**Assessment Rule**	**Low**	**High**	**p-value**
S1: Give/Give	0.9107	0.8929	(1, 1, 1, 1)	ALLC	0.0536	0.0238	0.142
S2: Keep/Give	0.2619	0.1964	(0, 0, 0, 0)	ALLD	0.0238	0.0476	0.329
S3: Give/Keep	0.8512	0.8274	(1, 0, 1, 0)	Image Scoring	0.5536	0.5833	0.589
S4: Keep/Keep	0.1548	0.1012	(1, 0, 1, 1)	Simple Standing	0.0595	0.0714	0.665
			(1, 0, 0, 1)	Stern Judging	0.0060	-	-
			(1, 0, 0, 0)	Shunning	0.0655	0.0714	0.846
			(1, 1, 1, 0)	Mild Shunning	0.1488	0.1131	0.366
			Other		0.0893	0.0893	1.000

Notes: *(Left section)* Numbers indicate the share of respondents that perceive *Give* to be seen as appropriate by the majority in response to the respective signal. *(Right section)* Numbers indicate the share of participants who selected each respective norm based on elicitation in the Interim Phase. Reported p-values are derived from binary probit models, where the dependent variable is an indicator for the chosen norm and the explanatory variable is an indicator for the cost condition with standard errors clustered at the session level. Number of observations = 168 per cost condition. The full set of norms is reported in Table A7 of the Appendix. Additional results from multinomial logit models that include control variables are presented in Table A8 of the Appendix.

5.4 Interim Phase - Group norms

For the group norm elicitation, subjects were asked to indicate the behavior they would recommend to the entire group (i.e., their session) in response to the signals “S3: Give/Keep” and “S4: Keep/Keep.” In the *Info* treatments, these recommendations were displayed before the beginning of Phase 2, whereas in the *NoInfo* treatments, they were shown only afterwards. However, since subjects did not know the timing of the disclosure when making their decisions, I continue to pool the treatments for this analysis. The results are presented in Table 7.

Table 7: **Group norms on individual level**

(1) Giving recommended			**(2) Norms**				
Signal	**Low**	**High**	**Norm**	**Assessment Rule**	**Low**	**High**	**p-value**
S3: Give/Keep	0.8869	0.8631	(1, 1)	Simple Standing	0.5536	0.5000	0.177
S4: Keep/Keep	0.5952	0.5238	(0, 0)	ALLD	0.0714	0.1131	0.173
			(1, 0)	Image Scoring	0.3333	0.3631	0.505
			(0, 1)	Stern Judging	0.0417	0.0238	0.285

Notes: *(Left section)* Numbers indicate the share of respondents that indicate *Give* as recommended behavior for the respective signal. *(Right section)* Numbers indicate the share of participants who selected each respective norm based on elicitation in the Interim Phase. Reported p-values are derived from binary probit models, where the dependent variable is an indicator for the chosen norm and the explanatory variable is an indicator for the cost condition with standard errors clustered at the session level. Number of observations = 168 per cost condition. Additional results from multinomial logit models that include control variables are presented in Appendix Table A9.

The outcomes of the group norm elicitation suggest that, under both cost conditions, the most commonly recommended behavior was to choose *Give* after both “S3: Give/Keep” and “S4: Keep/Keep”—that is, the pattern (1, 1), which corresponds to the Simple Standing norm. This behavior was endorsed by 55.36% of subjects under low costs and by 50.00% under high costs. While the difference is not significant, it nevertheless clearly runs counter to Hypothesis 3, which predicted a higher incidence of this norm when the cost of giving is high. There is also no evidence that subjects were more likely to recommend the behavior (0, 1), aligned with Stern Judging, under high costs. The share of recommendations consistent with this norm was very low across both cost conditions, with no significant differences observed. The norm (1, 0), related to Image Scoring, was endorsed by 33.33% and 36.31% of subjects, respectively—much lower than the shares observed for this behavior in terms of the perceived social injunctive norm. Again, there are no significant differences between cost conditions. Table A9 in the Appendix presents results from multinomial logit models that include additional control variables. The results suggest that participants who reported stronger tendencies toward negative (direct) reciprocity were less likely to recommend the norm associated with Simple Standing (1, 1) and more likely to favor the norm associated with Image Scoring (1, 0).

Figure 5 illustrates the relationship between participants' indicated personal injunctive norms and their group norm recommendations. In each panel, the left side displays the behavior that participants personally considered appropriate in response to the signals "S3: Give/Keep" and "S4: Keep/Keep," while the right side shows the behavior they recommended as a group norm. Connecting bars between the two sides indicate whether personal norms and recommendations align or diverge. Overall, in both cost conditions, participants tended to recommend for the group what they personally regarded as appropriate. A notable exception is a substantial subset of participants who personally endorsed (1, 0), related to Image Scoring, but recommended behavior (1, 1), aligned with Simple Standing, as the group norm.

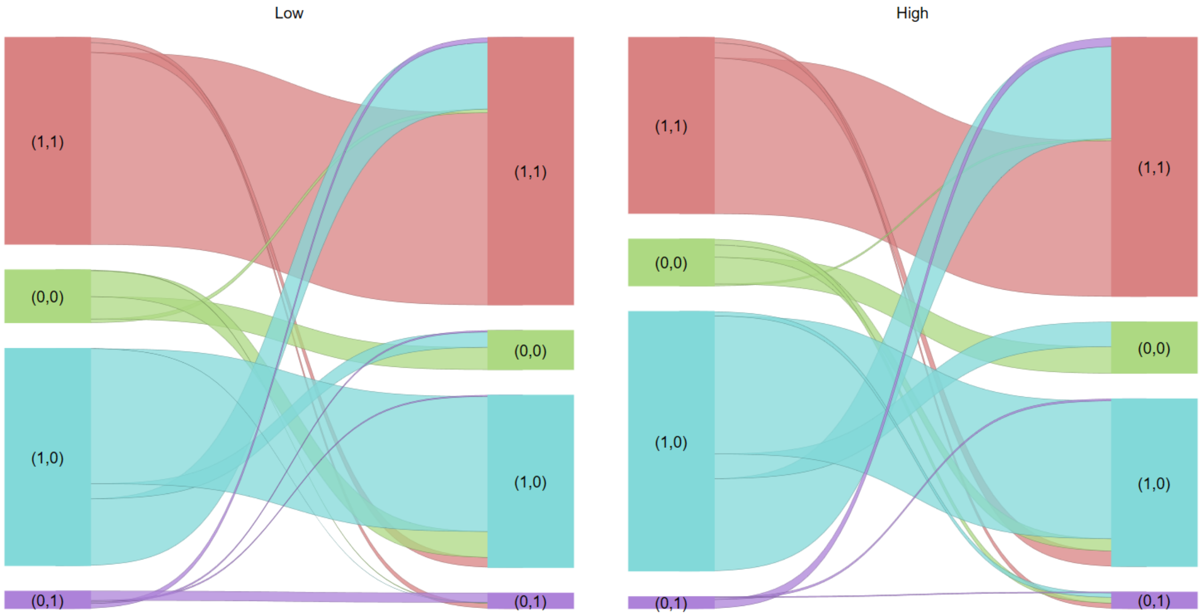


Figure 5: **Alignment between personal injunctive norms and group norm recommendations**

Notes: The Sankey diagram visualizes the relationship between participants' indicated personal injunctive norms (left side of each panel) and their group norm recommendations (right side of each panel) for signals "S3: Give/Keep" and "S4: Keep/Keep." The width of the flows represents the proportion of participants making each specific combination of judgments. The left panel displays responses under the low-cost conditions, while the right panel corresponds to the high-cost conditions.

Finally, before turning to the results from Phase 2, Table 8 summarizes the outcomes of the group norm elicitation at the session level. The table reports how many sessions arrived at each specific norm, based on a simple majority of participants' recommended behaviors. Since each session comprised 14 participants, ties were possible and were communicated to participants accordingly.

The results indicate that individual recommendations led to three distinct norms at the session level. The most frequently observed norm is $(1, 1)$, corresponding to Simple Standing, which was chosen in the majority of sessions and was particularly prevalent in the low-cost treatments. Under high-cost conditions, the norms corresponding to Simple Standing $(1, 1)$

and Image Scoring $(1, 0)$ appeared with nearly equal frequency. In both cost conditions, the norm $(0, 1)$, related to Stern Judging, was never selected at the session level. This distribution suggests substantial agreement among participants regarding the appropriate action following signal "S3: Give/Keep," but a lack of consensus concerning the appropriate response to signal "S4: Keep/Keep." This is further evidenced by sessions that produced the norm $(1, x)$, where a clear majority recommended *Give* for S3, but a tie between *Give* and *Keep* occurred for S4.

Table 8: **Group norms on session level**

Norm	**Assessment Rule**	**Low_NoInfo**	**Low_Info**	**High_NoInfo**	**High_Info**
(1, 1)	Simple Standing	4	5	3	2
(1, 0)	Image Scoring	1	1	1	3
(1, x)		1	-	2	1

Notes: Numbers indicate the number of sessions in which the respective norm was the outcome of the majority recommendation for the given signals. The notation $(1, x)$ indicates that the majority recommended *Give* for signal "S3: Give/Keep," while there was a tie between *Give* and *Keep* for signal "S4: Keep/Keep."

5.5 Phase 2 - Giving rates

I now turn to behavior in Phase 2, where variation in group norm information provision became relevant. Accordingly, the results are presented separately for the four treatments. I begin by summarizing overall giving rates before examining whether providing norm information increases alignment with the group norm. Figure 6 shows the giving rates by treatment for Phases 1 and 2. The results indicate that, across all four treatments, giving significantly increased in Phase 2 compared to Phase 1 ($p = 0.091$ for *Low_NoInfo*; $p < 0.001$ for *Low_Info*; $p = 0.003$ for *High_NoInfo*; and $p = 0.022$ for *High_Info*).[11] Giving rates

[11] Based on binary probit models with the decision (*Give* = 1, *Keep* = 0) as the dependent variable and an indicator for the phase (Phase 1 or Phase 2) as explanatory variable; standard errors clustered at the individual level. Random-effects panel probit models yield qualitatively similar findings.

conditional on signal are comparable to those found in Phase 1: subjects were more likely to give when their co-player chose to give in the previous round and more likely to choose *Keep* otherwise (see Table A10 in the Appendix).

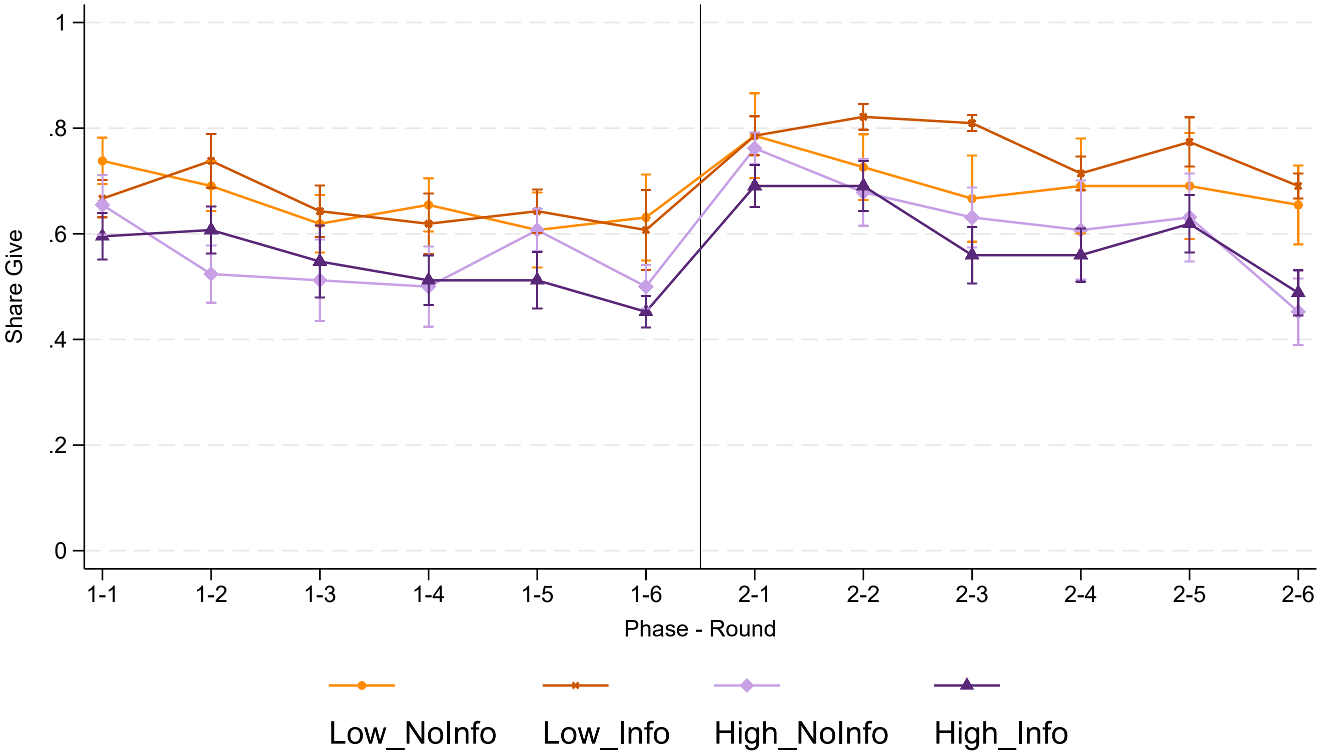


Figure 6: **Giving rates by treatment and signal (Phase 1 and Phase 2)**

Notes: Markers indicate the average share of *Give* by round of Phase 1 and Phase 2. For Round 5 of each phase (strategy method), the implemented choice was used. Light orange (circle) markers indicate *Low_ NoInfo*; dark orange (cross) markers indicate *Low_ Info*; light violet (diamond) markers indicate *High_ NoInfo*; dark violet (triangle) markers indicate *High_ Info*. Whiskers indicate standard errors clustered at the session level.

To assess whether the provision of group norm information increases giving, I estimated binary probit models separately for each cost condition, using data from both phases and all rounds with standard errors clustered at the individual level. The dependent variable is a binary indicator for whether a subject chose *Give*. The explanatory variables include an information treatment indicator (*NoInfo* or *Info*), an indicator for the phase (Phase 1 or Phase 2), and their interaction. Based on these models, I calculate second differences in the estimated probabilities of giving to compare changes in giving behavior from Phase 1 to Phase 2 across the two information conditions. Below, I report the p-values for the second differences; graphical illustrations of the estimated probabilities are provided in Figure A2 in the Appendix. For the low-cost treatments, the second difference indicates a larger increase (6.7 percentage points) in giving when group norm information was provided; however, this effect is only weakly statistically significant ($p = 0.070$). For the high-cost treatments, the second difference is negative (-1.4 percentage points), indicating a smaller increase in giving under *Info* compared to *NoInfo*; however, this difference is not statistically significant ($p = 0.713$).

As outlined above, not all sessions converged on the same group norm. Therefore, comparisons across treatments with different norms could be misleading. To address this, I also present estimates restricted to sessions in which participants indicated the same group norm information. Under low costs, four sessions in the *NoInfo* treatment and five sessions in the *Info* treatment converged on the same norm: (1, 1) related to Simple Standing. Applying the approach described above to only these sessions, the results again indicate a positive effect of information provision on giving under low costs; however, this effect is again only weakly statistically significant ($p = 0.074$). Under high costs, three sessions under *NoInfo* and two sessions under *Info* treatments selected the same norm: (1, 1). When comparing only these sessions, there is no significant difference ($p = 0.411$).

5.6 Phase 2 - Norm alignment

To investigate whether subjects were more likely to align their behavior with the group norm when this information was available, I first constructed indicator variables equal to 1 if a subject's behavior aligned with the group norm in a given session, and 0 otherwise.[12] These indicators were created separately for each signal ("S3: Give/Keep" and "S4: Keep/Keep") and for each elicitation method (direct response and strategy method). For example, if the group norm in a given session was (1, 0), the alignment indicator for S4 equals 1 if a subject chose *Keep* after S4, and 0 if the subject chose *Give*. These variables were constructed for both Phase 1 and Phase 2. To examine the effect of group norm information on alignment, I estimated binary probit models with standard errors clustered at the individual level, using the alignment indicator as the dependent variable. The explanatory variables included an indicator for the information condition (*NoInfo* or *Info*), an indicator for the phase (Phase 1 or Phase 2), and their interaction. Figure 7 shows results based on decisions made via the strategy method. The results indicate that subjects overall are somewhat more likely to align behavior with the group norm in Phase 2, with no evidence that the information provision about the group norm has a significant effect on alignment. In most cases, the second difference is close to zero. The only exception is the low-cost condition and signal S4, where the second difference is about 8.1 percentage points, though this difference is not statistically significant ($p = 0.304$). Likewise, in all other cases, no statistically significant differences were found ($p = 0.862$ for *Low* cost and S3, $p = 0.316$ for *High* cost and S3, and $p = 0.967$ for *High* cost and S4).

The results based on decision-making via direct elicitation do not yield statistically significant effects either. For signal "S3: Give/Keep," I find no evidence that second differences have an effect of information provision ($p = 0.954$ for *Low* and "S3: Give/Keep," $p = 0.684$ for *High* and "S3: Give/Keep"). Under high costs, alignment with the group norm even decreased after the norm was revealed before Phase 2. However, a similar decline is observed in the corresponding *NoInfo* condition, suggesting that the decrease cannot be attributed to the provision of group norm information. For signal "S4: Keep/Keep," there is likewise no indication of increased alignment when norm information was available ($p = 0.385$ for *Low*,

[12] For subjects in sessions where no majority was reached, the variable was coded as missing.

$p = 0.674$ for *High*). It should be noted, however, that the number of observations for S4 is relatively low under direct elicitation, as this signal appeared infrequently. The results are illustrated in Figure A3 in the Appendix. Taken together, these findings provide no support for Hypothesis 4, which stated that subjects would align their behavior with the group norm when this information is available. Overall, the provision of group norm information did not lead to greater alignment between individual behavior and the stated group norm.

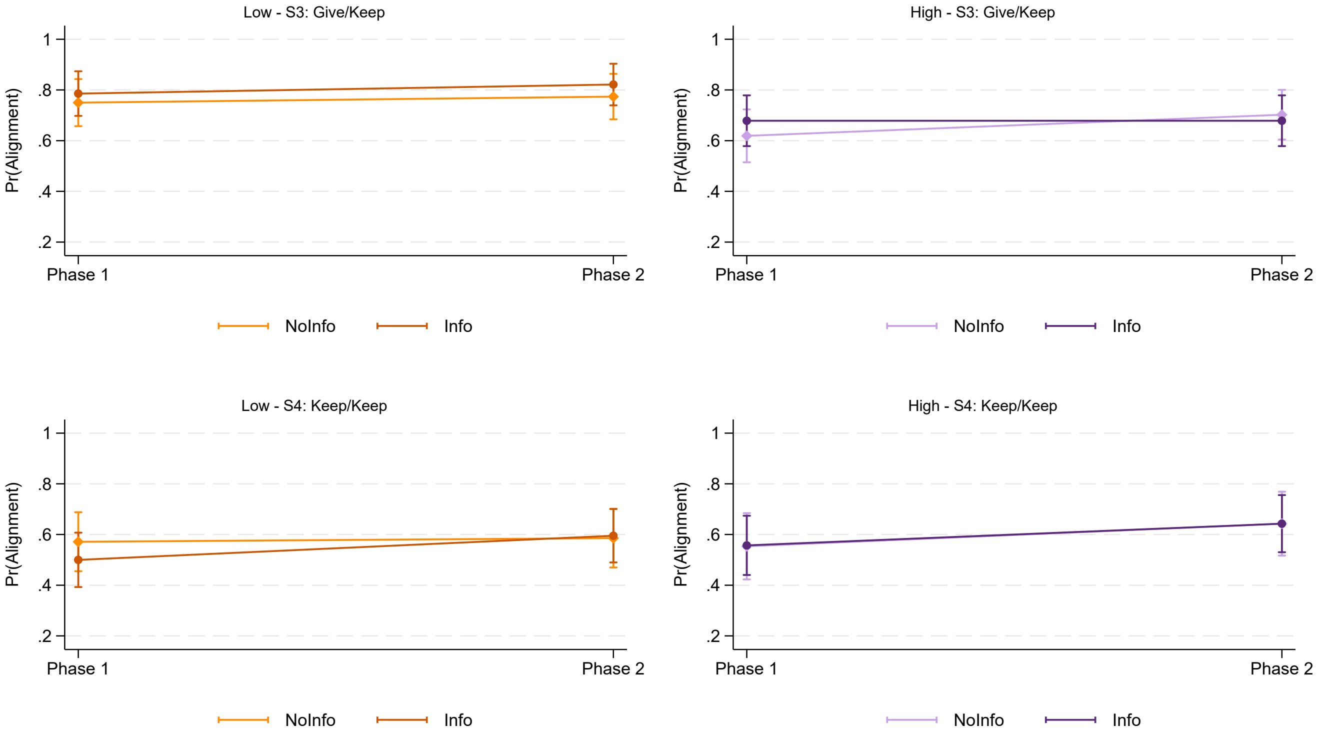


Figure 7: **Effects of norm information on alignment with group norms for strategy method**

Notes: Markers indicate estimated probabilities of norm-aligned behavior by phase (Phase 1 vs. Phase 2) and information condition (*NoInfo* vs. *Info*), based on binary probit models with standard errors clustered at the individual level. Whiskers indicate 95% confidence intervals. The top row shows results for alignment with the group norm following signal "S3: Give/Keep" under low and high cost conditions, respectively. The bottom row shows outcomes for "S4: Keep/Keep." Number of observations: 280 (top row left), 280 (top row right), 252 (bottom row left), 196 (bottom row right).

6 Discussion and Conclusion

This study aimed to deepen our understanding of the normative foundations of behavior under indirect reciprocity. While prior research has primarily focused on descriptive norms—documenting what people do—this work explored whether these behaviors reflect underlying personal beliefs and shared expectations about what is considered appropriate. Specifically, the study examined how individuals morally evaluate others based on past actions, drawing on three established assessment rules: Image Scoring, Simple Standing, and Stern Judging. At the heart of this inquiry lies a key question in moral evaluation: is cooperation inherently good and defection inherently bad, or do reputational judgments depend on context and justification? The study also investigated whether norm coordination at the group level—particularly when made salient—influences cooperative behavior and norm alignment. Since certain norms are theoretically better at sustaining long-term cooperation, and it is often assumed that groups adhere to shared norms, understanding how people converge on or diverge from these standards holds both theoretical and practical significance.

Consistent with previous work (e.g., Engelmann and Fischbacher 2009; Seinen and Schram 2006; Wedekind and Milinski 2000), the results demonstrate that indirect reciprocity can sustain substantial levels of cooperation, even under relatively unfavorable cost conditions. This suggests that the reputational mechanisms underpinning indirect reciprocity are robust across a range of contexts. Furthermore, to the best of my knowledge, this is only the second study to employ the strategy method in this domain, following Gaudeul et al. (2021). The alignment of findings across both studies—particularly under comparable cost conditions—reinforces the validity of their theoretical model and highlights the replicability of key behavioral patterns. With these broader insights in place, I now turn to the specific contributions of this study, which offer a more detailed understanding of the normative underpinnings and context sensitivity of behavior under different cost regimes.

I began by investigating whether notions of appropriate behavior depend on the difficulty of the cooperation problem. Specifically, I hypothesized that higher cooperation costs would lead participants to endorse norms that distinguish between justified and unjustified defection, reflecting increased sensitivity to contextual factors. However, I found no evidence supporting this hypothesis—neither in personal injunctive norms nor in perceived social injunctive norms. Rather, the findings suggest that participants' behavior reflects a shared normative foundation, with both personal beliefs and perceptions of others' expectations supporting the appropriateness of Image Scoring. In both cost conditions, the single most endorsed personal norm was Image Scoring. Even more striking, at least 50% of participants in both conditions believed that others regarded Image Scoring as the most appropriate norm. This suggests that behavior consistent with Image Scoring was not merely strategic but rooted in shared normative understandings of what is appropriate. This has important implications for interpreting giving behavior in this and similar experiments. For example, giving after the "S4: Keep/Keep" signal might be interpreted as viewing the earlier defection as justified, and thus as an endorsement of norms that differentiate between reasons for defection. However, participants who believe that most others follow Image Scoring may instead have given in such situations simply to be given to in the following interaction. This highlights how relying solely on descriptive behavior—without accounting for personal and perceived social injunctive norms—can result in misleading conclusions.

The strong endorsement of Image Scoring—both personally and socially—was accompanied by low support for norms that permit justified defection. This is reflected in participants' responses: depending on the cost condition, only about 44% or 34%, respectively, viewed giving after "S4: Keep/Keep" as appropriate. These findings are consistent with Yamamoto et al. (2020), who similarly observed limited support for the idea that people systematically evaluate such defections positively. Even more striking, only about 15% or 10%, respectively, of participants believed that most others considered such giving appropriate. This is somewhat surprising, given that conditional cooperation—rewarding cooperation and withholding help after defection—is well established in direct reciprocity. Yet in the context of indirect reciprocity, such conditional responses did not appear to be regarded as normatively appropriate, either personally or socially. While participants did not clearly endorse justified defection as a personal or perceived social norm, many nevertheless recommended it as the appropriate group norm—particularly those who had previously indicated preferences aligned with Image Scoring. This suggests that, although individuals may not personally view justified defection as appropriate, they may come to see it as strategically necessary or useful when asked to coordinate on a group norm. Such a shift highlights a distinction between private convictions and collective reasoning in norm formation. There is also modest evidence, under low-cost conditions, that making this group norm public may increase giving. Although this effect was only weakly significant, it suggests that publicly endorsing a group norm can influence cooperative behavior to some extent, even when that norm is not strongly held at the individual level.

Consistent with the findings of Gaudeul et al. (2021), Stern Judging did not appear to play a substantial role in terms of descriptive behavior. The data also suggests that most participants personally disapproved of withholding giving from someone who had previously given, regardless of the context. Less than 15% of participants across both cost conditions endorsed such behavior as personally appropriate, indicating broad rejection of the core principle underlying Stern Judging. This suggests that, despite its theoretical appeal, Stern Judging may not align with individuals' intuitive moral judgments or prevailing social expectations in the context of indirect reciprocity. This was further underscored by the fact that *Keep* was never recommended as appropriate behavior in response to the "S3: Give/Keep" signal on a session level. This is somewhat surprising, as the idea that differentiates Stern Judging from other norms—that one should not reward bad behavior—is a common moral intuition shared across many cultures. Yet this reasoning appeared to hold little normative weight in the context of the study.

With regard to the information provision variation, the results showed that while somewhat increased norm alignment was observed there was no significant difference between the information and no-information conditions. Except for a weakly significant effect on cooperation under low costs, information about the group norm had little impact on behavior. Notably, in all treatments, cooperation increased significantly in phase 2 compared to phase 1. This indicates that the mere engagement with norms during the interim phase may have impacted behavior, regardless of whether the outcomes of the elicitation were made public or not.

In summary, the findings suggest that participants' perceptions of appropriate behavior corresponded more closely to Image Scoring than to theoretically more effective norms such as Simple Standing and Stern Judging, particularly in terms of punishing unjustified defection and rewarding justified cooperation. This highlights two important avenues for future research: assessing whether Image Scoring indeed underperforms compared to these alternative norms in sustaining cooperation in social dilemmas with human subjects, and exploring why and under which circumstances participants are reluctant to distinguish between justified and unjustified defection and cooperation. Future research could also address two key limitations of this study to provide deeper insights into norms under indirect reciprocity. First, future studies could provide subjects with more comprehensive information on past behavior to unambiguously assess reputation according to norms beyond the first-order. One potential reason for the suggested dominance of Image Scoring might be that under the given information structure, this norm allows for clear judgments about a subject's reputation. Although the scenarios used to elicit injunctive norms also allowed for unambiguous higher-order assessments, participants may have remained anchored in the experience of the main task, where such information was not always available. As a result, they may have gravitated toward simpler norms like Image Scoring, which are easier to apply in information-poor environments. Second is the relatively low number of interactions. While this design ensured perfect stranger matching—i.e., pure indirect reciprocity—it may have reduced the need to coordinate on more robust, context-sensitive norms, making simpler heuristics like Image

Scoring appear sufficient. Future studies could address this by extending the number of interactions to better capture the strategic relevance of norms such as Simple Standing or Stern Judging. At the same time, more interactions would require rematching instead of perfect stranger matching for practical reasons, which could in turn dilute the distinction between direct and indirect reciprocity by increasing the chance of repeated encounters with the same individuals.

References

Alexander, Richard D. (1987). *The Biology of Moral Systems*. New York: Aldine de Gruyter.

Axelrod, Robert (2006). *The Evolution of Cooperation*. Revised edition. New York: Basic Books.

Bašić, Zoran and Elena Verrina (2024). “Personal norms — and not only social norms — shape economic behavior”. In: *Journal of Public Economics* 239, p. 105255. DOI: 10.1016/j.jpubeco.2024.105255.

Bolton, Gary E., Elena Katok, and Axel Ockenfels (2005). “Cooperation among strangers with limited information about reputation”. In: *Journal of Public Economics* 89.8, pp. 1457–1468. DOI: 10.1016/j.jpubeco.2004.03.008.

Brown, R. P. and A. Phillips (2005). “Letting bygones be bygones: Further evidence for the validity of the Tendency to Forgive scale”. In: *Personality and Individual Differences* 38.3, pp. 627–638. DOI: 10.1016/j.paid.2004.05.017.

Chen, Daniel L., Max Schonger, and Christopher Wickens (2016). “oTree—An open-source platform for laboratory, online, and field experiments”. In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97. DOI: 10.1016/j.jbef.2015.12.001.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde (2009). “Homo reciprocans: Survey evidence on behavioural outcomes”. In: *Economic Journal* 119.536, pp. 592–612. DOI: 10.1111/j.1468-0297.2008.02242.x.

Engelmann, Dirk and Urs Fischbacher (2009). “Indirect reciprocity and strategic reputation building in an experimental helping game”. In: *Games and Economic Behavior* 67.2, pp. 399–407. DOI: 10.1016/j.geb.2008.12.006.

Gaudeul, Antoine, Claudia Keser, and Sebastian Müller (2021). “The Evolution of Morals under Indirect Reciprocity”. In: *Games and Economic Behavior* 126, pp. 251–277. DOI: 10.1016/j.geb.2021.01.004.

Greiner, Bernd (2015). “Subject pool recruitment procedures: organizing experiments with ORSEE”. In: *Journal of the Economic Science Association* 1.1, pp. 114–125. DOI: 10.1007/s40881-015-0004-4.

Hilbe, Christoph, Lukas Schmid, Josef Tkadlec, and Martin A. Nowak (2018). “Indirect reciprocity with private, noisy, and incomplete information”. In: *Proceedings of the National Academy of Sciences* 115.48, pp. 12241–12246. DOI: 10.1073/pnas.1810565115.

Krupka, Erin L. and Roberto A. Weber (2013). “Identifying social norms using coordination games: Why does dictator game sharing vary?” In: *Journal of the European Economic Association* 11.3, pp. 495–524. DOI: 10.1111/jeea.12006.

Leimar, Olle and Peter Hammerstein (2001). “Evolution of cooperation through indirect reciprocity”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1468, pp. 745–753. DOI: 10.1098/rspb.2000.1573.

Milinski, Manfred, Dirk Semmann, T. C. M. Bakker, and Hans-Jurgen Krambeck (2001). “Cooperation through indirect reciprocity: Image scoring or standing strategy?” In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1448, pp. 2495–2501. DOI: 10.1098/rspb.2001.180.

Nowak, Martin A. (2006). “Five Rules for the Evolution of Cooperation”. In: *Science* 314.5805, pp. 1560–1563. DOI: 10.1126/science.1133755.

Nowak, Martin A. and Karl Sigmund (1998). “Evolution of indirect reciprocity by image scoring”. In: *Nature* 393, pp. 573–577. DOI: 10.1038/31225.

Ohtsuki, Hisashi and Yoh Iwasa (2004). “How should we define goodness?—Reputation dynamics in indirect reciprocity”. In: *Journal of Theoretical Biology* 231.1, pp. 107–120. DOI: 10.1016/j.jtbi.2004.06.005.

Pacheco, J. M., F. C. Santos, and F. A. C. C. Chalub (2006). “Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity”. In: *PLOS Computational Biology* 2.12, e178. DOI: 10.1371/journal.pcbi.0020178.

Rand, David G. and Martin A. Nowak (2013). “Human Cooperation”. In: *Trends in Cognitive Sciences* 17.8, pp. 413–425. DOI: 10.1016/j.tics.2013.06.003.

Santos, F. P., F. C. Santos, and J. M. Pacheco (2018). “Social norm complexity and past reputations in the evolution of cooperation”. In: *Nature* 555, pp. 242–245. DOI: 10.1038/nature25763.

Seinen, Ivar and Arthur Schram (2006). “Social status and group norms: Indirect reciprocity in a repeated helping experiment”. In: *European Economic Review* 50.3, pp. 581–602. DOI: 10.1016/j.euroecorev.2004.10.005.

Sugden, Robert (1986). *The Economics of Rights, Co-operation, and Welfare*. London: Palgrave Macmillan UK.

Swakman, Vera, Lucas Molleman, Aljaz Ule, and Martijn Egas (2016). “Reputation-based cooperation: Empirical evidence for behavioral strategies”. In: *Evolution and Human Behavior* 37.3, pp. 230–235. DOI: 10.1016/j.evolhumbehav.2015.12.001.

Trivers, Robert L. (1971). “The Evolution of Reciprocal Altruism”. In: *The Quarterly Review of Biology* 46.1, pp. 35–57. DOI: 10.1086/406755.

Wedekind, Claus and Manfred Milinski (2000). “Cooperation through image scoring in humans”. In: *Science* 288.5467, pp. 850–852. DOI: 10.1126/science.288.5467.850.

Yamamoto, H., T. Suzuki, and R. Umetani (2020). “Justified defection is neither justified nor unjustified in indirect reciprocity”. In: *PLOS ONE* 15.6, e0235137. DOI: 10.1371/journal.pone.0235137.

A Appendix

A.1 Example reputation information

Figure A1 shows an example of the reputation information a player received before decision-making (in this case: Round 4). In the experiment, the illustration was accompanied by the following text:

“**Explanation:** Your current co-player chose *Give* in Round 3 and knew at that time that the other player had previously chosen *Keep* in Round 2. In Round 2, this other player had information about the respective other player’s behavior in Round 1, but you do not have this information.”

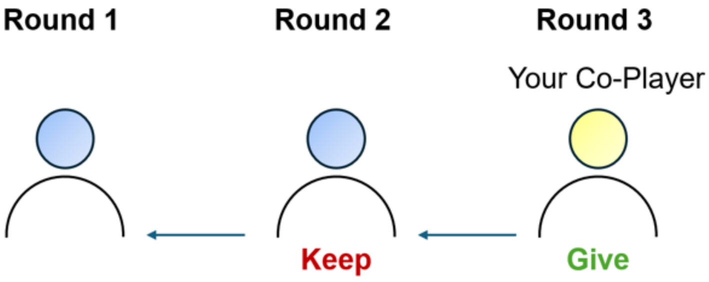


Figure A1: **Example for presentation of reputation**

Notes: In this example a player is in Round 4. Her current co-player chose *Give* in the previous round and knew that her receiver chose *Keep* in the round before that. Further information is not available.

A.2 Overview sample and sessions

Table A1: **Overview sample**

	All	**Low_NoInfo**	**Low_Info**	**High_NoInfo**	**High_Info**	**p-value**
Gender						
Female	0.56	0.64	0.55	0.54	0.52	0.386
Male	0.37	0.31	0.35	0.41	0.42	0.430
Divers	0.04	0.00	0.05	0.04	0.06	0.184
Prefer not to tell	0.03	0.05	0.06	0.02	0.00	0.136
Age						
Mean	26.14	26.57	26.02	25.45	26.51	0.207
Behavioral class						
Yes	0.26	0.18	0.26	0.30	0.29	0.282
Experiments						
None	0.39	0.40	0.35	0.46	0.35	0.329
1-2 Experiments	0.36	0.40	0.38	0.29	0.38	0.386
3-4 Experiments	0.16	0.11	0.18	0.18	0.19	0.445
5 or more	0.08	0.08	0.10	0.07	0.08	0.958
Forgiveness						
Mean	14.66	13.98	14.46	15.45	14.75	0.151
Positive Reciprocity						
Mean	18.61	18.94	18.51	18.56	18.43	0.723
Negative Reciprocity						
Mean	8.42	8.54	8.54	8.77	7.82	0.545
Difficulty						
Mean	2.10	2.14	2.37	2.06	1.85	0.222
Observations	336	84	84	84	84	-

Notes: The *p*-value in the last column indicates the probability from joint χ^2 or Kruskal–Wallis tests across all treatments. "Forgiveness" is based on the sum of four items from the "Tendency to Forgive Scale" (Brown and Phillips 2005; see S10 in the post-experiment survey). "Positive Reciprocity" and "Negative Reciprocity" are based on the sum of three items each (see S9 in the post-experiment survey), following, e.g., Dohmen et al. 2009. "Difficulty" refers to Question S1 in the post-experiment survey.

Table A2: **Overview of sessions**

Session	**Date**	**Treatment**	**Subjects**
Session 1	December 11, 2024	High_Info	14
Session 2	December 11, 2024	Low_Info	14
Session 3	December 17, 2024	High_NoInfo	14
Session 4	December 17, 2024	Low_NoInfo	14
Session 5	January 21, 2025	High_NoInfo	14
Session 6	January 23, 2025	High_Info	14
Session 7	January 23, 2025	Low_Info	14
Session 8	January 28, 2025	Low_NoInfo	14
Session 9	March 26, 2025	High_NoInfo	14
Session 10	March 27, 2025	Low_NoInfo	14
Session 11	April 2, 2025	High_Info	14
Session 12	April 2, 2025	Low_Info	14
Session 13	April 23, 2025	High_NoInfo	14
Session 14	April 29, 2025	Low_Info	14
Session 15	May 7, 2025	High_Info	14
Session 16	May 7, 2025	Low_NoInfo	14
Session 17	May 14, 2025	High_NoInfo	14
Session 18	May 15, 2025	Low_NoInfo	14
Session 19	May 21, 2025	High_Info	14
Session 20	May 28, 2025	Low_Info	14
Session 21	July 9, 2025	High_NoInfo	14
Session 22	July 10, 2025	Low_NoInfo	14
Session 23	July 16, 2025	High_Info	14
Session 24	July 17, 2025	Low_Info	14

Notes: The table shows all experimental sessions with dates, treatment conditions, and number of subjects per session.

A.3 Additional results: Phase 1 - Giving rates and personal descriptive norms

Table A3: **Giving by treatment and signal (Phase 1)**

	Low	**High**
Round 1		
None	70.24% (168)	62.50% (168)
Round 2		
Give	76.27% (118)	68.57% (105)
Keep	60.00% (50)	36.51% (63)
Average	71.43% (168)	56.55% (168)
Round 3-4 & 6		
Give/Give	71.32% (258)	64.71% (187)
Keep/Give	39.08% (87)	32.65% (98)
Give/Keep	78.08% (73)	61.54% (91)
Keep/Keep	48.84% (86)	35.16% (128)
Average	62.90% (504)	50.40% (504)
Round 5 (Strategy Method)		
Give/Give	72.02% (168)	65.48% (168)
Keep/Give	57.14% (168)	35.71% (168)
Give/Keep	76.79% (168)	64.88% (168)
Keep/Keep	48.21% (168)	40.48% (168)
Average	62.50% (168)	55.95% (168)
Overall Average	65.48% (1008)	54.37% (1008)

Notes: Numbers indicate percentage share of *Give* choices by treatment and signal, with number of observations in parentheses. The average for Round 5 (strategy method) is based on the implemented choice. The overall average is calculated using the implemented choice for Round 5.

Table A4: **Full personal descriptive norms (Phase 1)**

Norm	**Assessment Rule**	**Low**	**High**
(1, 1, 1, 1)	ALLC	0.2738	0.1667
(0, 0, 0, 0)	ALLD	0.0595	0.1250
(1, 0, 1, 0)	Image Scoring	0.1607	0.1607
(1, 0, 1, 1)	Simple Standing	0.0774	0.1429
(1, 0, 0, 1)	Stern Judging	0.0119	0.0060
(1, 0, 0, 0)	Shunning	0.0238	0.1012
(1, 1, 1, 0)	Mild Shunning	0.1250	0.0476
(1, 1, 0, 1)		0.0238	0.0119
(0, 1, 1, 1)		0.0119	0.0238
(1, 1, 0, 0)		0.0238	0.0179
(0, 1, 1, 0)		0.0357	0.0179
(0, 1, 0, 1)		0.0536	0.0179
(0, 0, 1, 1)		0.0179	0.0179
(0, 1, 0, 0)		0.0238	0.0536
(0, 0, 1, 0)		0.0655	0.0714
(0, 0, 0, 1)		0.0119	0.0179

Notes: Numbers indicate the share of participants who selected each respective norm based on elicitation using the strategy method in Round 5 of Phase 1. Number of observations = 168 per cost condition.

A.4 Additional results: Personal injunctive norms

Table A5: **Full personal injunctive norms**

Norm	Assessment Rule	Low	High
(1, 1, 1, 1)	ALLC	0.2500	0.1369
(0, 0, 0, 0)	ALLD	0.0238	0.0298
(1, 0, 1, 0)	Image Scoring	0.2917	0.4405
(1, 0, 1, 1)	Simple Standing	0.1369	0.1726
(1, 0, 0, 1)	Stern Judging	0.0179	0.0119
(1, 0, 0, 0)	Shunning	0.0476	0.0357
(1, 1, 1, 0)	Mild Shunning	0.1369	0.1012
(1, 1, 0, 1)		0.0060	0.0060
(0, 1, 1, 1)		0.0060	-
(1, 1, 0, 0)		0.0298	0.0179
(0, 1, 1, 0)		0.0060	-
(0, 1, 0, 1)		0.0600	0.0060
(0, 0, 1, 1)		0.0119	0.0060
(0, 1, 0, 0)		-	0.0119
(0, 0, 1, 0)		0.0238	0.0238
(0, 0, 0, 1)		0.0060	-

Notes: Numbers indicate the share of participants who selected each respective norm based on elicitation using the strategy method in the Interim Phase. Number of observations = 168 per cost condition.

Table A6: **Marginal effects at means from multinomial logit models, dependent variable: personal injunctive norm**

	(1)	(2)	(3)	(4)	(5)	(6)
	(1,1,1,1)	(0,0,0,0)	(1,0,1,0)	(1,0,1,1)	(1,0,0,1)	Other
High	-0.1250***	0.0019	0.1727***	0.0247	-0.0016	-0.0727
	(0.0442)	(0.0140)	(0.0507)	(0.0357)	(0.0047)	(0.0496)
Female	-0.0242	-0.0179	-0.0042	-0.0411	-0.0020	0.0894*
	(0.0412)	(0.0243)	(0.0625)	(0.0546)	(0.0028)	(0.0460)
Age	0.0077**	0.0006	-0.0039	-0.0077*	-0.0001	0.0033
	(0.0035)	(0.0015)	(0.0064)	(0.0045)	(0.0004)	(0.0046)
Game Theory	-0.0168	0.0136	0.0308	-0.0262	-0.0011	-0.0003
	(0.0443)	(0.0298)	(0.0534)	(0.0446)	(0.0043)	(0.0528)
Experiment (1-2)	-0.0663	0.0034	0.1324*	-0.0395	-0.0015	-0.0284
	(0.0523)	(0.0182)	(0.0736)	(0.0367)	(0.0080)	(0.0663)
Experiment (3-4)	-0.0820	0.0069	0.2894***	-0.0363	0.0027	-0.1806***
	(0.0627)	(0.0220)	(0.0600)	(0.0447)	(0.0177)	(0.0632)
Experiment (5 or more)	-0.0937	0.0089	0.1081	0.0391	-0.0120	-0.0504
	(0.0622)	(0.0248)	(0.0959)	(0.0749)	(0.0074)	(0.0720)
Forgiveness	0.0050	-0.0004	-0.0169***	0.0077	0.0004	0.0042
	(0.0035)	(0.0015)	(0.0060)	(0.0055)	(0.0003)	(0.0050)
Positive reciprocity	0.0129	-0.0030	0.0113	-0.0001	0.0009	-0.0220***
	(0.0095)	(0.0021)	(0.0084)	(0.0074)	(0.0010)	(0.0073)
Negative reciprocity	-0.0131***	-0.0018	0.0013	0.0042	0.0008	0.0087
	(0.0051)	(0.0022)	(0.0064)	(0.0056)	(0.0006)	(0.0061)
Observations	336					

Notes: Numbers indicate marginal effects at means (discrete probability effects for dummy variables) with standard errors clustered at the session level in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a nominal variable indicating the elicited personal injunctive norm: 1 = (1,1,1,1)/ALLC; 2 = (0,0,0,0)/ALLD; 3 = (1,0,1,0)/Image Scoring; 4 = (1,0,1,1)/Simple Standing; 5 = (1,0,0,1)/Stern Judging; 6 = Other. The base category for Female is Male, Diverse and Prefer not to tell. The base category for Experiment variables is Experiment (= 0).

A.5 Additional results: Perceived social injunctive norms

Table A7: **Full perceived social injunctive norms**

Norm	**Assessment Rule**	**Low**	**High**
(1, 1, 1, 1)	ALLC	0.0536	0.0238
(0, 0, 0, 0)	ALLD	0.0238	0.0476
(1, 0, 1, 0)	Image Scoring	0.5536	0.5833
(1, 0, 1, 1)	Simple Standing	0.0595	0.0714
(1, 0, 0, 1)	Stern Judging	0.0060	-
(1, 0, 0, 0)	Shunning	0.0655	0.0714
(1, 1, 1, 0)	Mild Shunning	0.1488	0.1131
(1, 1, 0, 1)		-	-
(0, 1, 1, 1)		0.0119	-
(1, 1, 0, 0)		0.0238	0.0298
(0, 1, 1, 0)		0.0119	0.0060
(0, 1, 0, 1)		0.0060	0.0060
(0, 0, 1, 1)		-	-
(0, 1, 0, 0)		0.0060	0.0179
(0, 0, 1, 0)		0.0119	0.0298
(0, 0, 0, 1)		0.0179	-

Notes: Numbers indicate the share of participants who selected each respective norm based on elicitation using the strategy method in the Interim Phase. Number of observations = 168 per cost condition.

Table A8: **Marginal effects at means from multinomial logit models, dependent variable: perceived social injunctive norms**

	(1)	(2)	(3)	(4)	(5)	(6)
	(1,1,1,1)	(0,0,0,0)	(1,0,1,0)	(1,0,1,1)	(1,0,0,1)	Other
High	-0.0081*	0.0019	0.0247	0.0102	-0.0000	-0.0287
	(0.0044)	(0.0018)	(0.0578)	(0.0233)	(0.0000)	(0.0602)
Female	-0.0043	0.0003	0.0118	-0.0185	0.0000	0.0106
	(0.0062)	(0.0008)	(0.0678)	(0.0339)	(0.0000)	(0.0481)
Age	-0.0002	0.0001	-0.0085*	-0.0001	0.0000	0.0087*
	(0.0006)	(0.0001)	(0.0049)	(0.0018)	(0.0000)	(0.0048)
Game Theory	0.0082	-0.0013	0.0394	-0.0114	0.0000	-0.0350
	(0.0071)	(0.0011)	(0.0541)	(0.0259)	(0.0000)	(0.0473)
Experiment (1-2)	0.0180	-0.0172	0.0419	-0.0329**	-0.0000	-0.0098
	(0.0209)	(0.0260)	(0.0782)	(0.0162)	(0.0000)	(0.0668)
Experiment (3-4)	0.0008	-0.0444**	0.1796**	-0.0372	0.0000	-0.0988
	(0.0197)	(0.0208)	(0.0712)	(0.0344)	(0.0000)	(0.0675)
Experiment (5 or more)	-0.0209*	-0.0186	0.0300	0.1061	-0.0000	-0.0966
	(0.0112)	(0.0265)	(0.0750)	(0.0693)	(0.0000)	(0.0888)
Forgiveness	0.0008*	-0.0000	-0.0064	0.0015	-0.0000	0.0041
	(0.0005)	(0.0002)	(0.0068)	(0.0027)	(0.0000)	(0.0069)
Positive reciprocity	0.0004	-0.0003	0.0063	0.0026	0.0000	-0.0091
	(0.0005)	(0.0002)	(0.0074)	(0.0050)	(0.0000)	(0.0075)
Negative reciprocity	-0.0005	-0.0002	-0.0021	-0.0046	-0.0000	0.0073
	(0.0007)	(0.0004)	(0.0081)	(0.0029)	(0.0000)	(0.0077)
Observations	336					

Notes: Numbers indicate marginal effects at means (discrete probability effects for dummy variables) with standard errors clustered at the session level in parentheses. Significance levels are indicated by: *p<0.10, **p<0.05, ***p<0.01. The dependent variable is a nominal variable indicating the elicited perceived social injunctive norm with 1 = (1,1,1,1)/ALLC; 2 = (0,0,0,0)/ALLD; 3 = (1,0,1,0)/Image Scoring; 4 = (1,0,1,1)/Simple Standing; 5 = (1,0,0,1)/Stern Judging; 6 = Other. The base category for Female is Male, Diverse and Prefer not to tell. The base category for Experiment variables is Experiment (= 0).

A.6 Additional results: Group norms

Table A9: **Marginal effects at means from multinomial logit models, dependent variable: group norm**

	(1)	(2)	(3)	(4)
	(1,1)	(0,0)	(1,0)	(0,1)
High	-0.0627	0.0368	0.0261	-0.0001
	(0.0410)	(0.0291)	(0.0418)	(0.0001)
Female	-0.0680	0.0295	0.0385	0.0000
	(0.0649)	(0.0282)	(0.0598)	(0.0001)
Age	0.0043	0.0011	-0.0054	-0.0000
	(0.0046)	(0.0022)	(0.0051)	(0.0000)
Game Theory	-0.1010	0.0124	0.0887	-0.0001
	(0.0638)	(0.0346)	(0.0609)	(0.0001)
Experiment (1-2)	-0.0185	0.0432	0.0268	-0.0515***
	(0.0675)	(0.0264)	(0.0685)	(0.0154)
Experiment (3-4)	-0.1955***	0.0531	0.1563*	-0.0139
	(0.0698)	(0.0487)	(0.0839)	(0.0297)
Experiment (5 or more)	-0.0868	0.0386	0.0641	-0.0158
	(0.0967)	(0.0399)	(0.1053)	(0.0414)
Forgiveness	-0.0020	0.0003	0.0017	0.0000
	(0.0066)	(0.0036)	(0.0058)	(0.0000)
Positive reciprocity	0.0105	-0.0143***	0.0039	-0.0000
	(0.0068)	(0.0040)	(0.0066)	(0.0000)
Negative reciprocity	-0.0248***	0.0044	0.0204***	0.0000
	(0.0065)	(0.0049)	(0.0059)	(0.0000)
Observations	336			

Notes: Numbers indicate marginal effects at means (discrete probability effects for dummy variables) with standard errors clustered at the session level in parentheses. Significance levels are indicated by: *p<0.10, **p<0.05, ***p<0.01. The dependent variable is a nominal variable indicating the elicited group norm with 1 = (1,1)/Simple Standing; 2 = (0,0)/ALLD; 3 = (1,0)/Image Scoring; 4 = (0,1)/Stern Judging. The base category for Female is Male, Diverse, and Prefer not to tell. The base category for Experiment variables is Experiment (= 0).

A.7 Additional results: Phase 2 - Giving rates and personal descriptive norms

Table A10: **Giving by treatment and signal (Phase 2)**

	Low_NoInfo	**Low_Info**	**High_NoInfo**	**High_Info**
Round 1				
None	78.57% (84)	78.57% (84)	76.19% (84)	69.05% (84)
Round 2				
Give	84.85% (66)	86.36% (66)	78.12% (64)	79.31% (58)
Keep	27.78% (18)	66.67% (18)	35.00% (20)	46.15% (26)
Average	72.62% (84)	82.14% (84)	67.86% (84)	69.05% (84)
Round 3-4 & 6				
Give/Give	77.48% (151)	74.39% (164)	74.26% (136)	70.73% (123)
Keep/Give	38.24% (34)	61.29% (31)	33.33% (36)	30.00% (40)
Give/Keep	83.33% (24)	86.84% (38)	51.85% (27)	44.12% (34)
Keep/Keep	44.19% (43)	63.16% (19)	28.30% (53)	38.18% (55)
Average	67.06% (252)	73.81% (252)	56.35% (252)	53.57% (252)
Round 5 (Strategy Method)				
Give/Give	67.86% (84)	76.19% (84)	71.43% (84)	59.52% (84)
Keep/Give	51.19% (84)	63.10% (84)	35.71% (84)	35.71% (84)
Give/Keep	75.00% (84)	78.57% (84)	61.90% (84)	67.86% (84)
Keep/Keep	48.81% (84)	47.62% (84)	41.67% (84)	39.29% (84)
Average	69.05% (84)	77.38% (84)	63.10% (84)	61.90% (84)
Overall Average	70.24% (504)	76.59% (504)	62.70% (504)	60.12% (504)

Notes: Numbers indicate percentage of *Give* choices by treatment and signal, with number of observations in parentheses. The average for Round 5 (strategy method) is based on the implemented choice. The overall average is calculated using the implemented choice for Round 5.

Table A11: **Full personal descriptive norms (Phase 2)**

Norm	**Assessment Rule**	**Low_NoInfo**	**Low_Info**	**High_NoInfo**	**High_Info**
(1, 1, 1, 1)	ALLC	0.2738	0.3810	0.1905	0.1667
(0, 0, 0, 0)	ALLD	0.1071	0.0595	0.1429	0.0952
(1, 0, 1, 0)	Image Scoring	0.1548	0.1786	0.2381	0.2738
(1, 0, 1, 1)	Simple Standing	0.1905	0.1071	0.1429	0.1310
(1, 0, 0, 1)	Stern Judging	0.0238	0.0119	0.0119	0.0476
(1, 0, 0, 0)	Shunning	0.0357	0.0476	0.0476	0.1071
(1, 1, 1, 0)	Mild Shunning	0.0833	0.0952	0.1071	0.0476
(1, 1, 0, 1)		0.0119	0.0119	0.0119	0.0238
(0, 1, 1, 1)		0.0119	0.0238	0.0119	-
(1, 1, 0, 0)		0.0357	0.0119	0.0238	0.0119
(0, 1, 1, 0)		0.0119	-	-	0.0119
(0, 1, 0, 1)		-	0.0119	0.0357	0.0119
(0, 0, 1, 1)		0.0119	0.0119	-	-
(0, 1, 0, 0)		-	0.0119	0.0119	0.0119
(0, 0, 1, 0)		0.0357	0.0238	0.0119	0.0476
(0, 0, 0, 1)		0.0119	0.0119	0.0119	0.0119

Notes: Numbers indicate the share of participants who selected each respective norm based on elicitation using the strategy method in Round 5 of Phase 2. Number of observations = 84 per treatment.

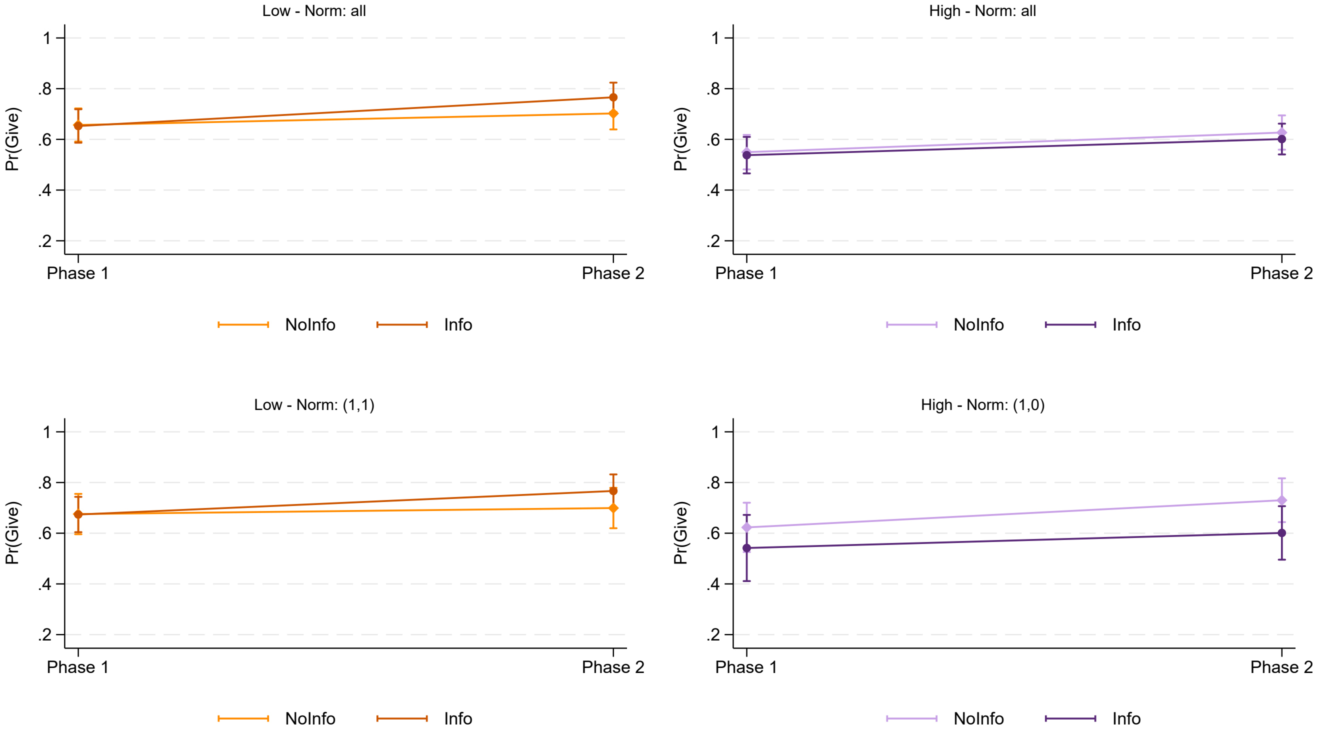


Figure A2: **Effects of group norm information on giving**

Notes: Markers indicate estimated probabilities of choosing *Give* by phase (Phase 1 or Phase 2) and information condition (*NoInfo* or *Info*) based on probit models with clustered standard errors at the individual level. Whiskers indicate 95% confidence intervals. The top row shows results for all sessions in the low-cost and high-cost treatments, respectively. The bottom row focuses on subsets of sessions that converged on a common group norm: sessions with norm $(1,1)$ (Simple Standing) under low costs, and sessions with norm $(1,0)$ (Image Scoring) under high costs. Number of observations: 2016 (top row left), 2016 (top row right), 1512 (bottom row left), 840 (bottom row right).

A.8 Additional results: Phase 2 - Norm alignment

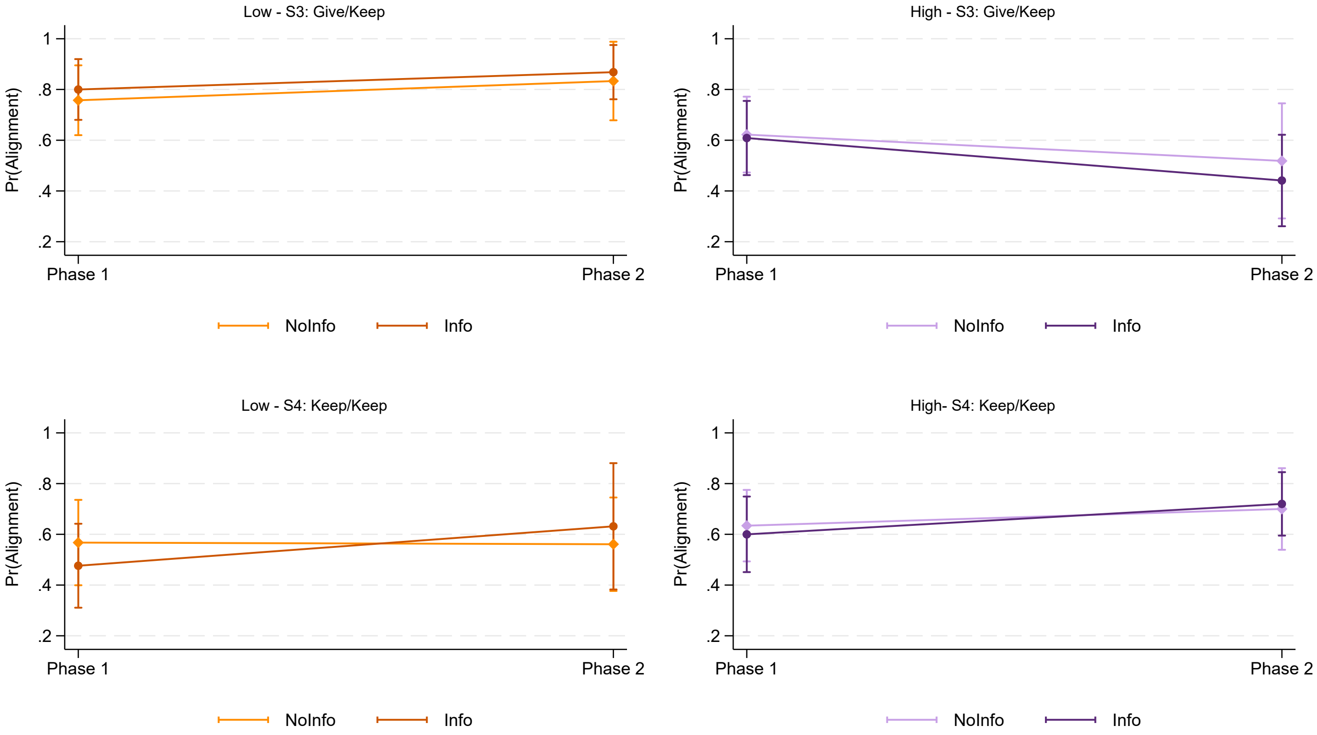


Figure A3: **Effects of norm information on alignment with group norms for direct elicitation**

Notes: Markers indicate estimated probabilities of norm-aligned behavior by phase (Phase 1 or Phase 2) and information condition (*NoInfo* or *Info*), based on probit models with standard errors clustered at the individual level. Whiskers indicate 95% confidence intervals. The top row shows results for alignment with the group norm following signal "S3: *Give/Keep*" under low- and high-cost conditions, respectively. The bottom row shows outcomes for "S4: *Keep/Keep*." Number of observations: 135 (top row left), 152 (top row right), 139 (bottom row left), 176 (bottom row right).

A.9 Instructions, quiz, decision screen

[The following instructions are for the high-cost treatments. The instructions for the low-cost treatments are identical, except for the cost parameter at various stages. Variations related to the info treatment do not affect these instructions and were communicated within the computer program.]

Welcome to the Experiment!

General Information

In this experiment, you will earn money. How much money you earn depends on your decisions and the decisions of your co-players. You only play with real players. These are the participants who are currently in the room with you. There are no simulated players (e.g., bots).

During the experiment, the payoffs are displayed in Lab Dollars (LD). The more LD you collect, the more money you will earn. The LDs are converted into euros (€) at the end of the experiment. For every LD you collect during the experiment, you will receive 1.5 cents as a payoff. In other words, you will receive 1.50€ for every 100 LD. For example, if you have collected 1200 LD at the end of the experiment, you will receive 18€. If the conversion does not result in a full amount, it will be rounded up to the next full 10 cent amount (e.g. 17.34€ to 17.40€). Payment is made in cash at the end of the experiment, without other participants being present.

The entire experiment will be conducted on the computer. Please avoid talking or communicating in any other way with other participants during the experiment. Please switch off your mobile phone.

We now come to the rules of today's experiment. This experiment is completely independent of other experiments you may have already participated in.

Rules of the Game

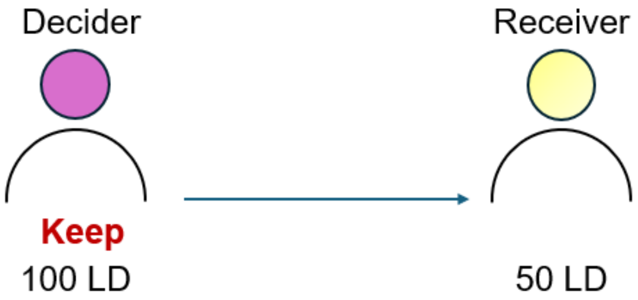
Games, Rounds, and Players

The experiment is divided into two game phases, and the rules of the game apply equally to both phases. Each game phase consists of six rounds. In each round, you will be assigned to a co-player, forming a group of two people. You will never form a group with the same co-player more than once.

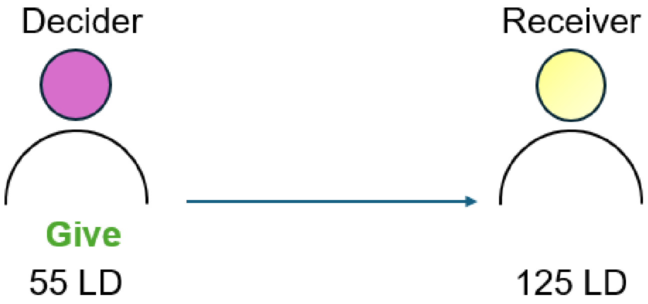
Decision and Roles

In each round, one participant will act as the decider and the other as the receiver. The decider chooses between “Keep” and “Give.” This decision will determine both the decider’s and the receiver’s payoffs. The payoffs for each option are as follows:

Keep: The decider receives 100 LD; the receiver receives 50 LD.



Give: The decider receives 55 LD; the receiver receives 125 LD.



Important

In each round, there is only one decider and one receiver per group. Nevertheless, both you and your co-player must first make the decision between "Keep" and "Give" in each round as if you were the decider. The computer then randomly determines which of you is actually the decider. The decision previously made by the person drawn then determines the payoffs for both of you. The decision made by the person not drawn has no relevance for the payoffs. Since you do not know at the time of the decision whether it will be relevant or not, you should always make your decision as if you were the decider and decide on the payoff for yourself and the payoff of your co-player.

Example 1

You choose “Keep” and your co-player chooses “Give.” The computer determines that you are the decider. This means that only your decision is relevant for the payoffs. → Payoffs: You receive 100 LD; your co-player receives 50 LD.

Example 2

You choose “Keep” and your co-player chooses “Give.” The computer determines that your co-player is the decider. This means that only your co-player’s decision is relevant for the payoffs. → Payoffs: You receive 125 LD; your co-player receives 55 LD.

Example 3

You choose “Give” and your co-player chooses “Keep.” The computer draws determines that you are the decider. This means that only your decision is relevant for the payoffs. → Payoffs: You receive 55 LD; your co-player receives 125 LD.

Example 4

You choose “Give” and your co-player chooses “Keep.” The computer determines that your co-player player is the decider. This means that only your co-player’s decision is relevant for the payoffs. → Payoffs: You receive 50 LD; your co-player receives 100 LD.

Information on Previous Behavior

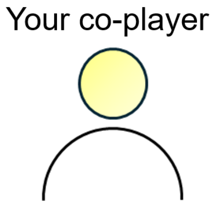
Before you make your decision, you will receive information about what your current co-player chose in the previous round. You will also get to know what information your co-player had about the other player in this round. Your current co-player will also receive this information about you.

Important

The information displayed is independent of whether or not your current co-player has actually been drawn by the computer as the decider. This also applies to the information about you.

Information in Round 1

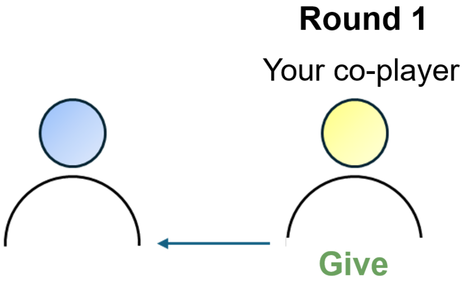
In the first round, you will not receive any information about your current co-player, as no decisions have yet been made. This will be displayed as follows:



Explanation: There is no information available about the behavior of your current co-player in a previous round.

Information in Round 2

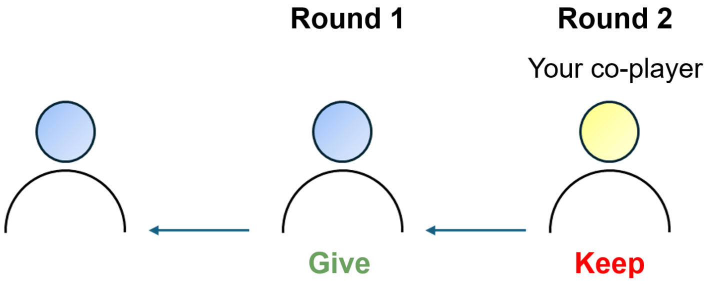
In Round 2, you will receive information about what your current co-player chose in the first round. This will be displayed, for example, as follows:



Explanation: Your current co-player chose **Give** in Round 1. In this Round 1, your co-player had no information about the behavior of the other player in a previous round.

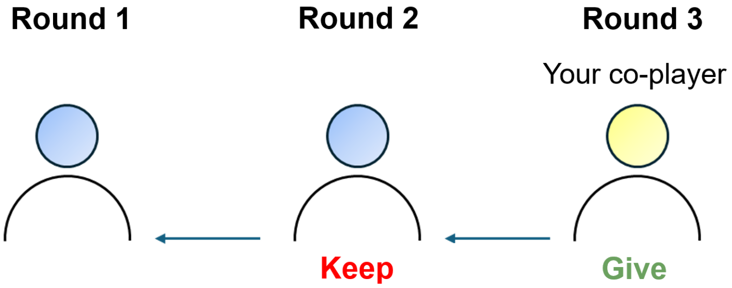
Information from Round 3 onwards

From Round 3 onwards, you will receive information about what your current co-player chose in the previous round and you will also find out what the player with whom your co-player played the previous round chose. This will be displayed to you in Round 3, for example, as follows:



Explanation: Your current co-player chose **Keep** in Round 2 and knew at this point that the other player had previously chosen **Give** in Round 1. This other player had no information in Round 1 about the behavior of their co-player in a previous round.

In Round 4 this is displayed, for example, as follows:



Explanation: Your current co-player chose **Give** in Round 3 and knew at this point that the other player had previously chosen **Keep** in Round 2. This other player had information in Round 2 about the behavior of their fellow player in Round 1. However, you do not have this information.

End of Phase 1

After Round 6, Phase 1 of the experiment ends.

Interim Phase

Between Phase 1 and Phase 2, there is an Interim Phase in which you are asked questions about the game. Here, you can earn additional payoffs. You will receive all necessary rules in the computer program. In the Interim Phase, the term *"the entire group of participants in today's experiment"* is sometimes used. This refers to all participants who are currently in the room. At the end of the Interim Phase, Phase 2 begins.

Phase 2

In Phase 2, the same rules apply as in Phase 1.

End of Phase 2

After Round 6, Phase 2 of the experiment ends. We will then ask you to complete a short questionnaire while we prepare the payments. Your total payoff consists of the sum of the payoffs from all rounds in the two Phases and the additional payoffs from the Interim Phase.

Quiz

1. Right or wrong? In each round, you are matched with a fellow participant. However, you never encounter the same participant twice. [Right]

   □ Right

   □ Wrong

2. Right or wrong? In each round, you and your co-player must both choose between "Keep" and "Give." The computer randomly determines whether your decision or your co-player's decision is relevant for the payouts. [Right]

   □ Right

   □ Wrong

3. Suppose you choose "Give" and your co-player chooses "Keep." The computer randomly selects you as the decider. What payoff do **you** receive for this round? [55 LD]

   □ 50 LD

   □ 55 LD

   □ 100 LD

   □ 125 LD

4. Suppose you choose "Keep" and your partner chooses "Give." The computer randomly selects you as the decider. What payoff does **your co-player** receive for this round? [50 LD]

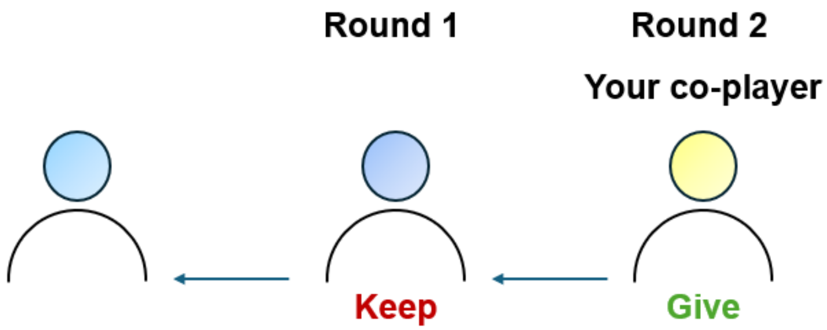
   □ 50 LD

   □ 55 LD

   □ 100 LD

   □ 125 LD

5. Suppose in round 3 you receive the following information about your co-player:



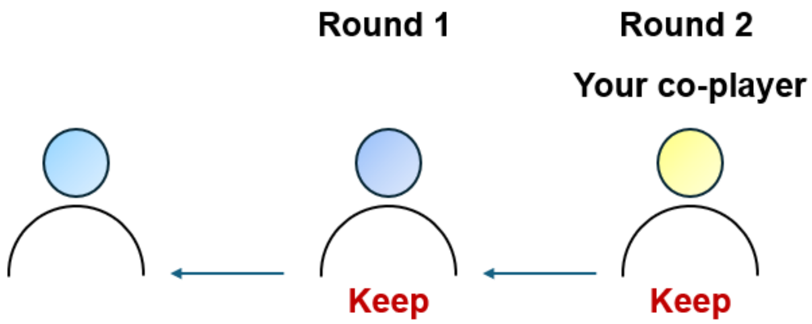
Which decision did your co-player make in round 2? [Give]

☐ Give

☐ Keep

☐ You do not have this information

6. Suppose in round 3 you receive the following information about your co-player:



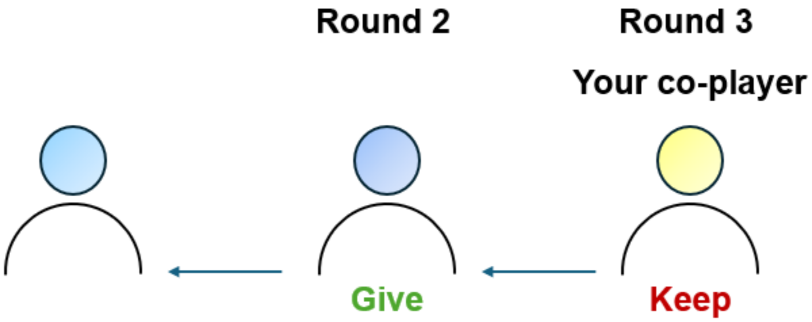
Which decision did your co-player make in round 1? [You do not have this information]

☐ Give

☐ Keep

- □ You do not have this information

7. Suppose in round 4 you receive the following information about your co-player:



Which decision did the player, with whom your co-player interacted in round 3, make in round 2? [Give]

- □ Give
- □ Keep
- □ You do not have this information

Decision Screen Example

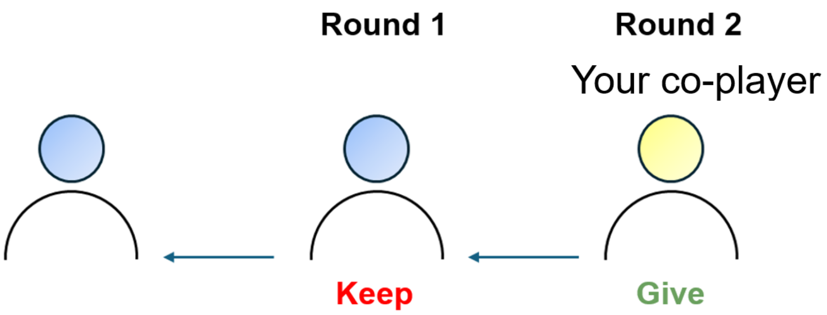
Round 3 – Decision

Your Choice: Choose between **Keep** and **Give**. Your decision will be displayed to your co-player in the next round.

Payoff:

- **Keep:** You receive 100 LD, your co-player receives 50 LD.
- **Give:** You receive 55 LD, your co-player receives 125 LD.

Information about your Co-Player:



Your current co-player chose **Give** in Round 2. The other player in this round chose **Keep** in Round 1. In Round 1, this other player had no information about their co-player's previous choices.

Please make your decision:

○ Keep ○ Give

Elicitation of Personal and Perceived Norms (Interim Phase)

Interim Phase 1: Personal Opinion on Game Situations

You will be presented with four situations that can occur in the game. In each, a decider (Person C) is in the third round and receives information about their co-player's (Person B) choice in round 2 when they played with Person A.

Payoffs associated with choices:

- **Keep:** Decider receives 100 LD, receiver 50 LD
- **Give:** Decider receives 55 LD, receiver 125 LD

Instructions: For each situation, indicate what you think the decider *should* do. After completing these questions for Signal 1, repeat the process for the remaining three signals.

Decision Screen

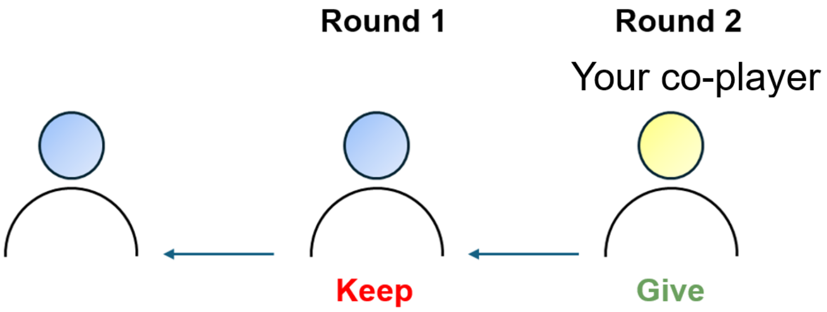
Round 3 - Decision

Decision: Choose between **Keep** and **Give**. Your decision will be displayed to your co-player in the next round.

Payoff:

- If you choose **Keep**, you receive 100 LD and your co-player receives 50 LD.
- If you choose **Give**, you receive 55 LD and your co-player receives 125 LD.

Information about Co-Player:



Your current co-player chose **Give** in Round 2. The other player in this round chose **Keep** in Round 1. In Round 1, this other player had no information about their co-player's previous behavior.

Please make your decision:

○ Keep ○ Give

Elicitation of Personal Injunctive, Perceived Social Injunctive, and Group Norms

[Example for the Interim Phase; participants repeat for all signals. High-cost treatment shown.]

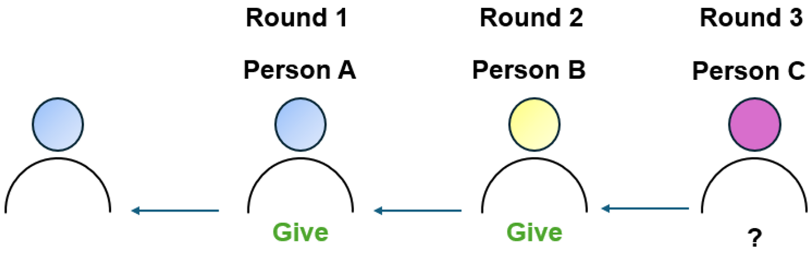
Interim Phase 1

Personal Opinion on Game Situations:

You will be presented with four situations that can occur in the game:

- In all situations, a decider (Person C) is in the third round and receives information about what their co-player (Person B) chose as decider in the second round when they played together with Person A. - Choice of the decider has the following payoffs:

- **Keep:** Decider receives 100 LD, receiver receives 50 LD.
- **Give:** Decider receives 55 LD, receiver receives 125 LD.



Explanation:

Person C plays together with person B in Round 3. Person B previously chose “Give” in Round 2. Person B played together with Person A in Round 2 and knew that Person A had previously chosen “Give” in Round 1 and at that time had no information about their co-player’s behavior from a previous round.

Personal Opinion Game Situation 1

This question is about your personal opinion regardless of the opinion of others. “Appropriate” behavior here is the behavior that you personally consider to be “right” or “moral.” It does not have to correspond to the behavior that you would actually choose or have chosen in this situation.

Please indicate what behavior you personally think is appropriate for Person C in this situation.

○ Keep ○ Give

Opinion of Others - Game Situation 1

Do you think that the majority of the participants in today's experiment have the same opinion as you?

○ Yes ○ No

Important: If you are correct with this assessment, you will receive 80 LD (= 1.20€) as an additional payment (see examples).

Example 1: If the majority has selected "Keep," you will receive 80 LD if your personal opinion is "Keep" and you have indicated "Yes" or if your personal opinion is "Give" and you have indicated "No."

Example 2: If the majority has selected "Give," you will receive 80 LD if your personal opinion is "Give" and you have indicated "Yes" or if your personal opinion is "Keep" and you have indicated "No."

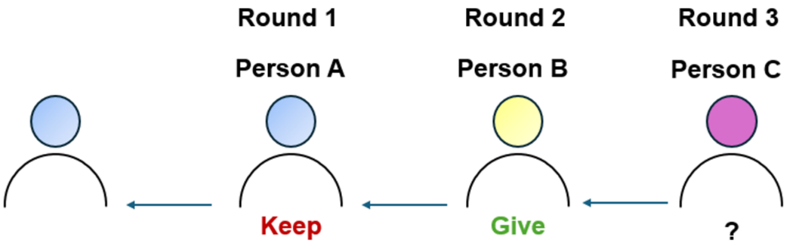
Elicitation Group Norms

[The following is an example for the elicitation of group norms in the Interim Phase for Signal 3. After answering this questions, participants answered exactly the same question for Signal 4.]

Recommended Behavior - Game Situation 3

Game Situation 3 is presented to you again below.

Game Situation 3



Explanation: Person C plays together with Person B in Round 3. Person B previously chose "Give" in Round 2. Person B played together with Person A in Round 2 and knew that Person A had previously chosen "Keep" in Round 1 and at that time had no information about their co-player's behavior from a previous round.

Opinions

There are different opinions about what behavior is appropriate for Person C in this situation. Two opinions are presented below.

Opinion 1: Person C should choose "Give." This is appropriate because Person B has previously chosen "Give" themselves.

Opinion 2: Person C should choose “Keep.” Although Person B has previously chosen “Give,” it would have been appropriate to choose “Keep” instead because Person A had previously chosen “Keep.”

Recommended Behavior

For this situation, a behavioral recommendation should now be created for the entire group of participants. This behavioral recommendation should help to develop a common understanding of what the participants should choose in this situation.

What behavior would you recommend for the entire group in this situation?

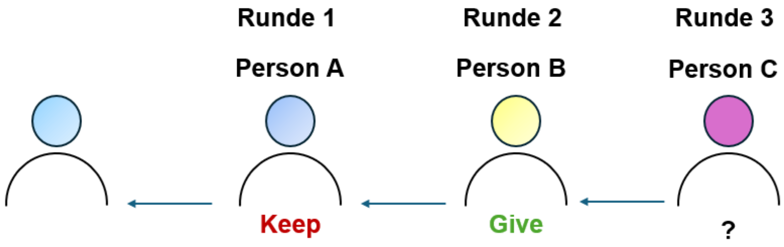
Important: All participants in today’s experiment answer this question. Before the start or after the end of the next game phase, the behavior recommended by the majority is announced for all participants in today’s experiment. Your individual recommendation will not be made public.

○ Keep ○ Give

Info screen

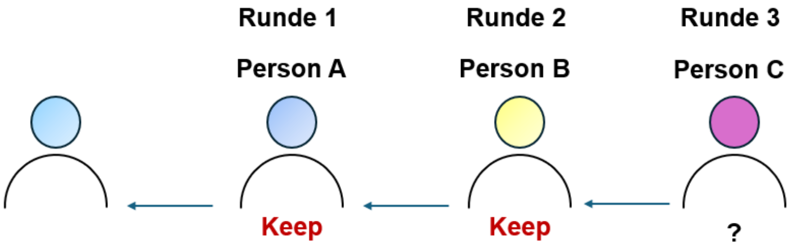
[Example for Simple Standing (1, 1)]

Non-binding Behavioral Recommendation – Game Situation 3



The majority of participants indicate "Give" as the non-binding behavioral recommendation for this situation.

Non-binding Behavioral Recommendation – Game Situation 4



The majority of participants indicate "Give" as the non-binding behavioral recommendation for this situation.

A.10 Post-experiment survey

S1. **Difficulty**

To what extent does the following statement apply to your participation in today's experiment?

It was difficult for me to understand what I had to do in this experiment.

□ (1) Does not apply at all

□ (2)

□ (3)

□ (4)

□ (5)

□ (6)

□ (7) Fully applies

S2. **Strategy – Game Phase 1**

Now think about game phase 1. Did you follow a specific strategy when deciding between "Give" and "Keep" in game phase 1? Please briefly describe your strategy:

S3. **Strategy – Game Phases 2**

Now think about game phase 2. Did you follow a specific strategy when deciding between "Give" and "Keep" in this phase? Please briefly describe your strategy:

S4. **Information**

Do you think the information provided to you was sufficient to make your decisions? If not, what else would you have liked to know before making your decisions?

S5. **Gender**

Please indicate your gender:

□ Male

□ Female

□ Diverse

□ Prefer not to say

S6. **Year of Birth**

Please select your year of birth:

□ 1900–2024 (dropdown)

S7. **Game Theory and Behavioral Economics**

Have you already attended a course that covered content from game theory and/or behavioral economics?

□ Yes

□ No

S8. **Experiment Experience**

How many economic experiments have you previously participated in? (This includes online experiments.)

□ 0

□ 1–2

□ 3–4

□ 5 or more

S9. **Reciprocity**

To what extent do the following statements apply to you personally?

□ If someone does me a favor, I'm willing to return it. (1–7)

□ I put in extra effort to help someone who has helped me before. (1–7)

□ I am willing to incur costs to help someone who has helped me in the past. (1–7)

□ If I suffer a serious injustice, I will make sure to take revenge at the next opportunity. (1–7)

□ If someone puts me in a difficult position, I will do the same to them. (1–7)

□ If someone insults me, I will behave insultingly toward them. (1–7)

S10. **Forgiveness**

To what extent do the following statements apply to you personally?

□ I get over emotional injuries relatively easily. (1–7)

□ When someone wrongs me, I often think about it for a long time. (1–7)

□ I tend to hold grudges. (1–7)

□ When others wrong me, I try to forgive and forget. (1–7)

S11. **Additional Comments**

Do you have any additional comments about the experiment you would like to share with us?