Reciprocity: On the Relative Importance and Interaction of Intention and Outcome Effects*

SIMON DATO† AND TIM FRIEHE‡

February 28, 2025

Outcomes and perceived intentions influence how individuals reciprocate others' actions. Using an experiment to disentangle the impact of outcomes and intentions, this paper provides first evidence about the relative importance and interaction of these determinants of reciprocal behavior. In our data, outcomes' impact on reciprocal behavior dominates that of intentions. Furthermore, intentions and outcomes interact in inducing reciprocity: the impact of a good outcome on reciprocal behavior is magnified by kind intentions. To uncover what drives outcome and intention effects, we elicit social norms from third parties and provide evidence of their significant explanatory power. However, comparing a social-preference-based explanation to the social-norm-based explanation reveals that the former explains choices better than the latter.

Keywords: Reciprocity; Intentions; Social Norms; Social Preferences; Experiment
JEL Codes: D63; D91; J16

*Funded by Germany's Excellence Strategy – EXC 2126/1 – 390838866.

†*Corresponding author:* EBS University for Business and Law, Rheingaustr. 1, D-65375 Oestrich-Winkel, Germany, E-mail address: simon.dato@ebs.edu

‡University of Marburg, Am Plan 2, D-35037 Marburg, E-mail address: tim.friehe@uni-marburg.de. Declarations of interest: none.

1 INTRODUCTION

Consider the following scenario: a senior consultant is searching for a junior consultant to join her team. The senior's manager determines whether recruitment involves minor or extensive screening, where the latter is more likely to identify high-ability candidates but comes at a higher cost for the manager. Hiring a high-ability candidate *directly* benefits only the senior consultant. However, the manager's willingness to invest in search and the search result (the junior's ability) will probably influence the senior's motivation and job performance. This would be aligned with evidence suggesting that outcomes and perceived kindness play significant roles in shaping how individuals reciprocate others' actions (e.g., Falk *et al.*, 2003, Falk *et al.*, 2008, McCabe *et al.*, 2003).

Despite the substantial evidence on the influence of outcomes and intentions on reciprocal behavior, little is known about their relative importance and interaction. Is the senior's motivation affected more by the manager's intention when choosing the kind of screening or by the ultimate outcome (i.e., the junior's ability)? Furthermore, does the senior's job-performance reaction to recruiting a high instead of a low-ability candidate (i.e., the outcome effect) depend on the manager's intention; that is, do intentions and outcomes interact in their impact on reciprocal behavior?

These questions are practically relevant in various settings. Knowledge about the relative importance and interaction of intention and outcome effects is crucial for the manager's optimal screening decision in our leading example as it helps to predict the senior's reciprocal behavior. For instance, a software developer might incur a private cost to carefully assess the needs of a client, aiming to increase the likelihood that the final product enhances the client's profitability. If the developer's effort is observable but not verifiable, the client can reciprocate by granting or blocking access to follow-up projects but cannot sue the developer for inadequate care. Similarly, an upstream firm might invest in precautions to improve the probability of timely delivery of inputs to its downstream customer, knowing that punctual delivery is also possible but less likely with low investment. In response, the downstream customer may reciprocate by adjusting their business dealings with the input provider.

Our paper contributes to the literature in two ways. First, by designing an experiment that isolates intention and outcome effects, we assess (a) whether outcome effects are more important than intention effects, (b) whether a norm focus can influence the relative importance of outcome effects – following Krupka and Weber (2009) – on either intentions, outcomes, or both, and (c) whether outcome and intention effects interact in their influence on reciprocal behavior. Second, we explore what drives intention and outcome effects as determinants of reciprocal behavior, examining the explanatory power of social norms (elicited following Krupka and Weber, 2013) and social preferences as described by Charness and Rabin (2002).

We use a modification of the moonlighting game (Abbink *et al.*, 2000). The first-mover (FM) chooses between two probability distribution over two fixed interim outcomes, one being preferred by the FM and the other being preferred by the second-mover (SM). The FM's choice that makes the SM's preferred outcome more likely represents kind intentions, while the FM's other choice renders the non-preferred outcome more likely and represents unkind intentions. After observing the FM's intention and the realized outcome, the SM rewards or punishes the FM. Our design separates intention and outcome effects by varying outcomes while holding intentions constant and vice versa, which is crucial for addressing our research question.[1]

Our data reveals that intentions and outcomes influence reciprocal behavior. SMs reward more or punish less when the FM shows kind instead of unkind intentions, even when it does not lead to a more preferred outcome. Likewise, after a given FM choice, SMs reward more or punish less when the preferred outcome is drawn instead of the non-preferred outcome. This suggests that the senior consultant would consider the manager's screening decision and the selected candidate's ability when assessing her job motivation.

After establishing the relevance of intentions and outcomes, we investigate their relative importance. The impact of average outcomes is more than twice that of average intentions. This suggests that altering outcomes influences reciprocal behavior more than modifying intentions. Consistently, the outcome effect dominates the intention effect when the two effects are opposing: SMs tend to reward their FM when the outcome is positive, but the intentions are unkind and to punish them when the outcome is negative, but the intentions are kind.

The relative importance of outcome effects remains unchanged when participants are provided with a norm focus regarding intentions, outcomes, or both, following Krupka and Weber (2009). While these different decision-making setups do not yield significant differences in reciprocal behavior, we find that the induced norm focus can influence SMs' evaluation of what constitutes socially appropriate behavior, similar to Gächter *et al.* (2017), who reported results from an experiment where only norms differed across treatments.

Having established the robustness of the relative importance of outcome effects, we demonstrate that intention and outcome effects interact. The average SM response to kind instead of unkind intentions (i.e., the intention effect) is magnified if the preferred outcome applies, while the response to the preferred instead of the non-preferred outcome (i.e., the outcome effect) is stronger when kind intentions apply. Intentions and outcomes are thus *complementary* in inducing positive reciprocity.

We conducted a norms experiment with a different set of subjects to investigate the drivers of

[1] The original moonlighting game and many other paradigms employed in previous papers on reciprocity featured a direct link between kind actions and high payoffs that hindered identification. While some papers have considered the possibility that *one* outcome can be reached via different intentions (e.g., Charness and Levine, 2007), our design allows this possibility for all outcomes.

intention and outcome effects, following Krupka and Weber (2013). Subjects evaluated SM's response options for all possible combinations of intentions and outcomes. We find that these ratings react to intention and outcome changes like choices: conditional on the preferred instead of the non-preferred outcome, rewards are rated more, and punishments less appropriate. Evaluating SM's alternatives conditional on kind instead of unkind intentions renders punishments significantly less appropriate without affecting the social appropriateness of rewards. Accordingly, a draw of the preferred outcome and kind intentions shifts the social norm toward positive reciprocity. Our norms data shows greater relative importance of outcomes (like our choice data) but no interaction between the outcome and the intention effects. When linking data from our choice and norms experiments, we find evidence for the explanatory power of social norms. A substantial part of the variation in reciprocal behavior may be due to a desire to comply with social norms. This is consistent with contributions such as Krupka and Weber (2013) and Kimbrough and Vostroknutov (2016).

However, in line with Gächter *et al.* (2013), social preferences à la Charness and Rabin (2002) provide a better fit for our choice data. Our results indicate that SMs are motivated by inequity aversion. They reduce (dis)advantageous payoff-inequity by rewarding (punishing) the FM. Strikingly, the norms coefficient becomes *negative* in a model combining social norms and preferences. This is because the social norm aligns with social efficiency concerns and is inconsistent with inequity aversion. When facing disadvantageous payoff inequity, the SM chooses costly punishment, which contrasts with the social efficiency motive inherent in social norms. In this regard, our paper contributes to the literature about the existence of a social norm of punishment (e.g., Fehr and Schurtenberger, 2018): when punishment is costly and serves only re-distributional purposes, it is not socially appropriate.[2]

The structure of the paper is as follows. We discuss the related literature in Section 2. In Section 3, we explain the experimental design and procedures. Section 4 provides a theoretical analysis of SM's choice across scenarios. Section 5 reports our empirical findings regarding choice data and social norms. Section 6 concludes.

2 LITERATURE

Our paper studies the relative importance and interaction of outcomes and intentions in shaping reciprocal behavior. While the acknowledgment of peoples' distributional concerns (e.g., Fehr and Schmidt, 1999) is widespread, fewer contributions account for intentions. Approaches to identify intention effects typically fall into one of two categories: (i) comparing responses to another party's choice to responses to an action implemented by a random draw (e.g., Offerman,

[2]This is maybe even more interesting when noting that the punishment in a one-shot interaction may arise out of a deterrence motivation because people anticipate spillovers to other contexts (e.g., Crockett *et al.*, 2014).

2002; Charness, 2004; Falk *et al.*, 2008), and (ii) comparing responses at a particular node in a game depending upon how it was reached (e.g., Falk *et al.*, 2003; McCabe *et al.*, 2003). Our design builds on Falk *et al.* (2008), who utilize the first approach to isolate the pure outcome fairness concern in an *intention-free* version of the moonlighting game.

In our design, outcomes and intentions are always at play. A given outcome can be reached via kind and unkind intentions, akin to the setup by Charness and Levine (2007). In their setting, principals first select a high or a low wage, and both levels can be brought to an intermediate level by a random move. Thus, *one* outcome level (the intermediate wage) is compatible with kind and unkind intentions on the principal's part. Charness and Levine (2007) demonstrate that a worker's effort determination is influenced by whether the principal's wage offer was high or low, showing that workers respond to the principal's intention.

Outcomes determined by actions and luck are relevant in the labor market, where effort is often unverifiable, and random factors can distort the verifiable and contractually used outcome. Rubin and Sheremeta (2016) explore a three-stage game where the principal determines the wage and desired effort level, the agent responds with effort, and the principal assigns a reward or punishment. They compare a treatment in which the outcome equals the effort level with treatments in which the outcome results from effort and a random draw. In these cases, the principal in the third stage may know how the outcome dis-aggregates into effort and luck. Principals assess their response conditional on effort and luck if they have the information, signifying that outcomes and perceived intentions matter for the principal's responses.[3] In our study, we start from the premise that both factors are relevant and aim to understand their relative importance and interaction precisely.

Friedrichsen *et al.* (2022) use a principal-agent setup to study whether SMs choose to be ignorant about the FM's intention to exploit wiggle room. In their design, FMs choose between low and high investment levels, affecting the success probability of a project whose proceeds are evenly split. SMs can then transfer points to their FM based on the project's success. In the treatment in which SMs would have to incur a symbolic cost to learn their FM's intention, many choose not to do so.[4] In our design, SMs state their response conditional on their FM's action and the outcome, eliminating observability issues, as this aspect is essential for our study of the relative importance and interaction.

We expect that SMs' evaluation of their FM's choice depends on the realized outcome, connecting our study to the literature on outcome bias (e.g., Baron and Hershey, 1988). Brownback

[3]The results by Rubin and Sheremeta (2016) were largely reproduced in Davis *et al.* (2017). In contrast, Kerschbamer and Oexl (2023) studies the three-stage game in a long-term principal-agent relationship and finds that the detrimental effect from the random shock is less pronounced.

[4]Friehe and Utikal (2018), Toussaert (2017) and Chan and Wolk (2023) contribute settings where the FM's intentions may remain hidden for other reasons.

and Kuhn (2019) study the outcome bias in a principal-agent setting where the agent can provide effort to raise the probability that the principal obtains a monetary gain, and the principal can punish the agent after payoffs are realized. They show that the principal's punishment of low effort is weaker when a gain materializes. In our study of the interaction, we investigate whether the outcome bias for kind intentions differs from that for unkind intentions, revealing results that contrast with previous findings.

To comprehend the drivers of SMs' choices, we follow Krupka and Weber (2013) and subsequent contributions by exploring the role of social norms. Additionally, we assess the explanatory power of social preferences using the setup introduced by Charness and Rabin (2002). Our paper aligns with Gächter *et al.* (2013), who consider a three-person gift-exchange game where the employer assigns wages to two employees choosing effort in sequence. They find that social norms do not predict the employees' effort choices when social norms and social preferences enter their empirical model. In our results, social norms even have a significant and *negative* coefficient in the most comprehensive model.

Our paper is also related to the literature on social dilemmas in which decision-makers' actions are implemented only with some probability (e.g., Rand *et al.*, 2015, Xiao and Kunreuther, 2016). For instance, Rand *et al.* (2015) consider a repeated interaction and argue that parties condition behavior on players' intentions when they are observable. In contrast, our paper considers a one-shot interaction, ensuring that considerations about a future interaction with the same participant do not contaminate the SM's choice.

3 Design

We implemented a modified "moonlighting game", as introduced by Abbink *et al.* (2000), a widely adopted paradigm to study positive and negative reciprocity (among others, see Cox *et al.*, 2008; Falk *et al.*, 2008; Van der Weele *et al.*, 2014; Cohn *et al.*, 2015). The one-shot game involved two randomly matched players acting sequentially. Both players received an endowment of 12 points to be used during the game.

Stage 1: First-Mover's Choice of Probability Distribution In the original "moonlighting game", FM selects an action a from the choice set $\{-6, -5, ..., 5, 6\}$, where any $a < 0$ signifies that FM gains a points while SM loses a points and any $a \geq 0$ signifies that FM loses a points while SM gains $3a$ points. We need to break the link between A's choice and the interim outcome to disentangle intentions from outcomes. We modify FM's choice so that she does not directly choose the interim outcomes but can allocate probability between outcomes.

We select two different interim outcomes from the moonlighting game: $a = -2$ (corresponding to an interim outcome $(\pi_1^{FM}, \pi_1^{SM}) = (14, 10)$) and $a = 2$ (corresponding to

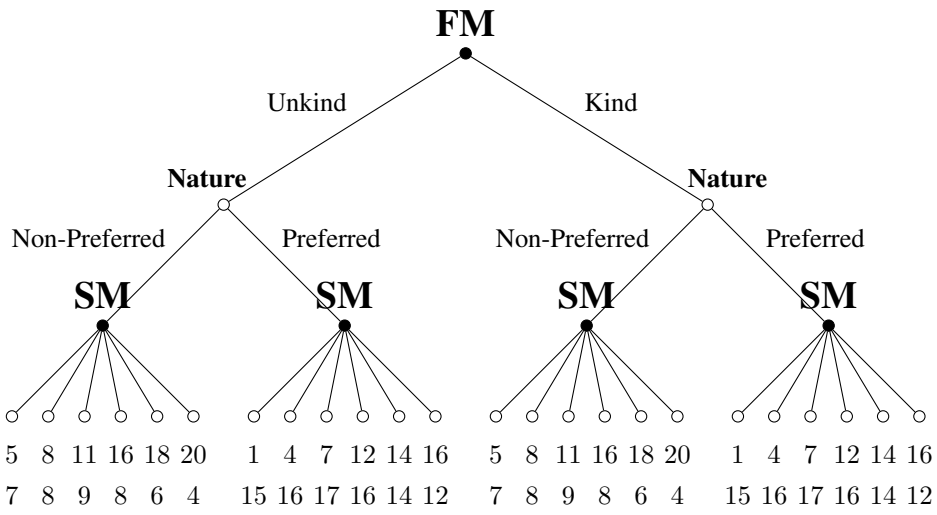


Figure 1: *Game Tree (Top Final Payoff for FM, Bottom Final Payoff for SM)*

$(\pi_2^{FM}, \pi_2^{SM}) = (10, 18)$). We will refer to interim outcome 2 as the *preferred* (abbreviated *p*) outcome from SM's point of view, as it yields higher own and total payoff as compared to the *non-preferred* (abbreviated *np*) outcome 1 for SM.[5] FM can choose a probability distribution under which the *non-preferred* outcome materializes with 80 percent probability. We will refer to this choice as the FM showing *unkind intentions* (abbreviated *ui*). Alternatively, FM can choose a probability distribution under which the *preferred* outcome materializes with 80 percent probability. We will refer to this choice as the FM showing *kind intentions* (abbreviated *ki*).

Stage 2: Second-Mover's Reciprocal Behavior After learning about FM's choice and the draw of the interim outcome, SM can choose action $b \in \{-3, -2, -1, 2, 4, 6\}$. SM can reward FM by choosing $b > 0$ and, hence, transferring b own points to FM. Alternatively, SM can punish FM by choosing $b < 0$. In that case, SM invests $|b|$ points to reduce FM's payoff by $3|b|$ points. SM's choice set is structured so that, regardless of the draw of the interim outcome, she can reduce, eliminate, and even reverse the payoff inequality present in the interim outcome. The final payoffs for FM and SM are given by

$$\Pi_i^{FM}(b) = \begin{cases} \pi_i^{FM} + b & \text{if } b > 0 \\ \pi_i^{FM} - 3|b| & \text{if } b < 0 \end{cases} \quad \text{and} \quad \Pi_i^{SM}(b) = \pi_i^{SM} - |b|, \quad i \in \{np, p\},$$

where the index i indicates whether the preferred or non-preferred outcome materialized. The game is summarized in Figure 1.

[5]Instructions were neutral. Please see the Supplementary Material for a translated version.

Treatments To assess the robustness of the relative importance and interaction of the intention and outcome effects, we conducted four treatments. By providing a norm focus, we aimed to shift attention to or away from outcomes. First, the neutral treatment, labeled BASE, serves as a baseline. In the treatment labeled INT, we provided a norm focus regarding intentions. We highlighted the impact of FM's choice on the probability distribution over interim outcomes and asked participants about the socially appropriate FM choice before letting them select their action b. This procedure is conceptually similar to that used in Krupka and Weber (2009). In treatment OUT, we aimed to draw attention to the interim *outcomes* by highlighting the positive impact of a draw of the preferred outcome for (i) SMs and (ii) total payoffs. Additionally, we again asked SMs about the socially desirable outcome before letting them select their action b. Finally, in treatment INTOUT, both aspects were made salient in this way.

Questionnaire We collected information on the subjects' economic preferences (risk, patience, reciprocity) using the items from Falk *et al.* (2018), their social value orientation (Murphy *et al.*, 2011) to understand how they value others' payoffs, their norm orientation (Bizer *et al.*, 2014) to learn about the extent to which players let themselves be guided by perceived norms, and their willingness to enforce norms using items from Traxler and Winter (2012) to contextualize the readiness with which SMs sanction their FM for non-compliance with norms after the experiment. Finally, participants completed a demographic survey.

Procedures The choice experiment was conducted online with the BonnEconLab, using hroot (Bock *et al.*, 2014) for online recruitment and the experimental software oTree (Chen *et al.*, 2016). Sessions lasted about one hour. After excluding subjects with more than two incorrect answers to 16 control questions, we have observations for 549 subjects (64% female). Subjects received a 4 Euro show-up fee and were compensated using an exchange rate of 1 point to 75 Euro Cent. Average earnings amount to 12.27 Euro.

We employed the strategy method to elicit SMs' choices. Each SM had to choose a level of b for every combination of FM choice and draw of the interim outcome. The chosen option for both the (i) actually chosen probability distribution over outcomes and (ii) drawn outcome was implemented and payoff-relevant.

Elicitation of Social Norms We elicited social norms from third parties following Krupka and Weber (2013). This process involved asking subjects to rate the choice options of the FM and the SM, i.e., the six possible levels of b for each potential scenario, along with the two choice options of FM. We utilized the scale from Chang *et al.* (2019): "very socially appropriate", "socially appropriate", "somewhat socially appropriate", "somewhat socially inappropriate", "socially inappropriate", and "very socially inappropriate". One appropriateness rating was

randomly selected, and a subject's rating on that item was compared to the session's mode. In addition to the 4 Euro show-up fee, a subject earned 20 Euro if their rating coincided with the most frequent rating of the other subjects. This method incentivized participants to reveal their perception of what is commonly regarded as socially (in)appropriate behavior rather than stating their private evaluation. For this experiment, we also relied on subjects from the BonnEconLab pool, the online recruitment hroot (Bock *et al.*, 2014), and the experimental software oTree (Chen *et al.*, 2016). After excluding subjects with more than two incorrect answers to 16 control questions, we have observations for 65 subjects (60% female). Average earnings amount to 12.65 Euro.

4 Theoretical Analysis

To provide a theoretical analysis of SM behavior, we rely on the setup by Charness and Rabin (2002). SM's utility function in scenario (i, j), where $i \in \{p, np\}$ and $j \in \{ki, ui\}$, reads

$$U_{i,j} = \Pi_i^{SM} + \rho \min\{\Pi_i^{FM} - \Pi_i^{SM}, 0\} + \sigma \max\{\Pi_i^{FM} - \Pi_i^{SM}, 0\} + \theta q_{\text{Selfish}} \left(\Pi_i^{SM} - \Pi_i^{FM}\right), \tag{1}$$

where $q_{\text{Selfish}} = 1$ if $j = ui$ (i.e., FM showed *unkind intentions*) and zero otherwise.

SMs' utility depends on their material payoff Π_i^{SM} and the difference between their and their FM's material payoff. The second term is relevant in case of *advantageous inequity*, i.e., when SM's material payoff is higher than FM's. Advantageous inequity increases SM's utility when $\rho < 0$ holds and decreases it otherwise. The third term differs from zero if FM's material payoff exceeds SM's, i.e., when *disadvantageous inequity* applies. Disadvantageous inequity decreases SM's utility when $\sigma < 0$ holds and increases it otherwise. The attitude towards and importance of (dis)advantageous payoff inequity is captured by the level of ρ (σ). Like Charness and Rabin (2002), we assume $\rho \geq \sigma$, so SM's preference for gains relative to FM is stronger when being behind than when being ahead. Likewise, we assume $\sigma \geq -1$ and $\rho \leq 1$, implying that inequity concerns do not outweigh material payoffs. *Competitive preferences* arise with $\sigma \leq \rho \leq 0$, *inequity aversion* by assuming $\sigma < 0 < \rho < 1$, and *social-welfare preferences* with $0 < \sigma \leq \rho \leq 1$ (Charness and Rabin, 2002). The fourth term incorporates intentions-based reciprocity via $\theta \geq 0$: if FM has *misbehaved*, which is assumed to be the case if they revealed *unkind intentions*, SM is more competitive towards FM.

As in Falk *et al.* (2008), we consider that SM chooses between different *changes* in FM's points, defined as $\Delta(i, j)$, where $i \in \{np, p\}$ and $j \in \{ki, ui\}$. $\Delta(\cdot)$ ranges from -9 (strongest punishment) to 6 (highest reward). The alternatives impact SM's material payoff independently of the scenario, so any conditionality of Δ on the scenario must stem from the remaining terms in SM's utility function. SM chooses starting from one of two interim outcomes, the non-preferred one with payoffs $(\pi_{np}^{FM}, \pi_{np}^{SM}) = (14, 10)$ or the preferred one with payoffs

$(\pi_p^{FM}, \pi_p^{SM}) = (10, 18)$. The possible levels of Δ can induce either advantageous or disadvantageous inequity. Conditional on the non-preferred outcome, disadvantageous inequity results at $\Delta \in \{-3, ..., 6\}$, and advantageous inequity when $\Delta = -9$. Conditional on the preferred outcome, disadvantageous inequity results when $\Delta = 6$, and advantageous inequity when $\Delta \in \{-9, ..., 2\}$. In addition, SM can induce payoff equality with $\Delta = -6$ ($\Delta = 4$) starting from the (non-)preferred outcome. SM's utility from choosing a specific level of Δ in scenario (i, j) is shown in Table 1.

Δ	$U_{p,ki}$	$U_{p,ui}$	$U_{np,ki}$	$U_{np,ui}$
-9	$15 - 14\rho$	$15 - 14\rho + 14\theta$	$7 - 2\rho$	$7 - 2\rho + 2\theta$
-6	$16 - 12\rho$	$16 - 12\rho + 12\theta$	8	8
-3	$17 - 10\rho$	$17 - 10\rho + 10\theta$	$9 + 2\sigma$	$9 + 2\sigma - 2\theta$
2	$16 - 4\rho$	$16 - 4\rho + 4\theta$	$8 + 8\sigma$	$8 + 8\sigma - 8\theta$
4	14	14	$6 + 12\sigma$	$6 + 12\sigma - 12\theta$
6	$12 + 4\sigma$	$12 + 4\sigma - 4\theta$	$4 + 16\sigma$	$4 + 16\sigma - 16\theta$

Table 1: SM's Utility For Different Δ in Scenario (i, j)

We are interested in how *unkind* instead of *kind intentions* and the *non-preferred* instead of the *preferred outcome* influence SM's choice of Δ. Regarding the former, we define the *intention effect*

$$\mathcal{I}(i) = \Delta^*(i, ki) - \Delta^*(i, ui),$$

as SM's behavioral adjustment when reacting to unkind instead of kind intentions for outcome i. In analogy, we define the *outcome effect*

$$\mathcal{O}(j) = \Delta^*(p, j) - \Delta^*(np, j),$$

as SM's change in behavior when choosing Δ conditional on the non-preferred instead of the preferred outcome for intention j.

When reacting to unkind instead of kind intentions, SMs become more competitive towards their FM. Formally, the term $\theta\left(\Pi_i^{SM} - \Pi_i^{FM}\right)$ is added to SM's utility function. This term makes tolerating disadvantageous inequity more costly and attaining advantageous inequity more desirable for SM. The utility-level change is higher for levels of Δ further away from the one inducing payoff equity, which we define as $\hat{\Delta}(i)$ where $\hat{\Delta}(p) = 4$ and $\hat{\Delta}(np) = -6$. Any behavioral adjustment will thus consist of choosing a lower Δ than with kind intentions:

Hypothesis 1. *SMs choose a (weakly) lower Δ when reacting to unkind instead of kind intentions. This produces a (weakly) positive intention effect $\mathcal{I}(i) \geq 0$, $i = np, p$.*

Concerning the outcome effect, Table 1 shows that SM tolerates disadvantageous inequity at many levels of Δ after a draw of the non-preferred outcome, and advantageous inequity at many

levels of Δ when the preferred outcome was drawn, where σ (ρ) is the parameter associated with (dis)advantageous inequity. For example, choosing $\Delta = 2$ instead of $\Delta = -3$ means that SM's utility changes by 6ρ after a draw of the preferred outcome and by 6σ after a draw of the non-preferred outcome. As $\rho \geq \sigma$, SM is more tolerant of FM's payoff, i.e., is less competitive or more altruistic, when advantageous inequity applies. Accordingly, choosing a lower level of Δ changes SM's utility more favorably when the non-preferred outcome was drawn. This leads to:

Hypothesis 2. *SMs choose a (weakly) lower Δ after a draw of the non-preferred outcome than after a draw of the preferred one. This produces a (weakly) positive outcome effect $\mathcal{O}(j) \geq 0$, $j = ki, ui$.*

Next, we explore how intentions and outcomes interact in their influence on SM's decisions. We define the interaction as the difference between the outcome effect conditional on kind intentions and the one conditional on unkind intentions, as

$$\mathcal{M} = \mathcal{O}(ki) - \mathcal{O}(ui).$$

By construction, it holds that $\mathcal{O}(ki) - \mathcal{O}(ui) = \mathcal{I}(p) - \mathcal{I}(np)$.

A first observation regarding the interaction is that it will be type-specific. For example, a very small intentions-based reciprocity parameter θ often means that the outcome effect is independent of intentions, so $\mathcal{M} = 0$ holds in many cases. In contrast, a high θ makes punishment levels dominate rewards and thereby will often cause $\mathcal{O}(ui) = 0$, which can induce considerable interaction $\mathcal{M}$ when SM reacts strongly to the variation in the interim outcome after kind intentions. For example, for competitive SMs with $\sigma < -\frac{1}{2} < \rho < \theta - \frac{1}{2}$ and inequity-averse SMs with $\rho > \frac{1}{3}$, we find $\mathcal{M} \geq 0$ when θ is sufficiently high. However, $\mathcal{M} < 0$ results for some types; for example, when $\mathcal{O}(ki) = 0$ and $\mathcal{O}(ki) > 0$.

A second observation is that the presence of the unkind intentions term $\theta(\Pi_i^{SM} - \Pi_i^{FM})$ does not directly change the tendency to lower Δ after the draw of the non-preferable outcome for an individual with fixed parameters. For example, choosing $\Delta = 2$ instead of $\Delta = -3$ means that SM's utility changes by $6\rho(-6\theta)$ after a draw of the preferred outcome and by $6\sigma(-6\theta)$ after a draw of the non-preferred outcome with (un)kind intentions. In other words, the outcome effect is not directly reinforced or counteracted by unkind intentions in the present framework as θ is outcome-independent. The discrete utility changes are symmetrically affected when moving from one level of Δ to another. However, it is crucial to note that unkind intentions introduce a tendency toward lower Δ independent of the draw of the outcome. This may mean that there is less room for an adjustment of Δ to the non-preferable outcome, implying that SM's outcome effect may be more likely muted when FM shows unkind intentions.

While we cannot derive general predictions at the individual level, we can analyze the interaction of intentions and outcomes at the aggregate level of SM types who differ in their ρ

and σ. For any $\theta > 0$, more SM types exhibit an outcome effect after kind intentions than after unkind intentions. Going beyond this binary categorization, we find that the level of the expected outcome effect is larger after kind than after unkind intentions, assuming that types are uniformly distributed over the ρ-σ space. Moreover, because the expected outcome effect after unkind intentions decreases in θ, we find that the expected $\mathcal{M}$ increases when SMs are more intentions-based reciprocal. Given these insights, we expect a positive interaction of kind intentions and a draw of the preferred outcome in shaping SM's choices.

Hypothesis 3. *The outcome effect after kind intentions (weakly) exceeds the one after unkind intentions,* $\mathcal{M} \geq 0$.

The theoretical analysis does not provide unambiguous predictions regarding the relative importance of intentions and outcomes for SM's choice. Generically, the relative importance varies with the parameters ρ and σ (determining outcome-based preferences) as well as θ (determining intention-based preferences). First, the intention effect (for both outcome draws) increases in θ, while the outcome effect after unkind (kind) intentions decreases (remains constant). Accordingly, an increase in the degree of intentions-based reciprocity shifts the relative importance towards intentions. Second, we analyze the impact of the outcome based preference parameters. Ceteris paribus, the higher ρ is, the stronger is the outcome effect (no matter FM intentions). Likewise, the higher σ is, the lower is the outcome effect. The effect of ρ and σ on the intention effect is less clear. For instance, if $\rho < -\frac{1}{2}$, the intention effect is zero but can be positive if $\rho \geq -\frac{1}{2}$: the intention effect cannot be monotonically decreasing in ρ. As the mathematical derivations have shown, however, an increase in ρ can also eliminate the intention effect. As an example, if $-\frac{1}{2} \leq \rho < \frac{1}{6}$, $\mathcal{I}(p) = 0 \Leftrightarrow \rho \geq \theta - \frac{1}{2}$. This demonstrates that the impact of ρ on the intention effect is non-monotonic. Similar arguments prove the same result for σ. Accordingly, the relative importance of intentions and outcomes is an empirical question and we abstain from formulating a hypothesis.

5 Results

We will first analyze SMs' choices regarding the relative importance and interaction of outcomes and intentions in Section 5.1. Next, we briefly analyze FMs' choices in Section 5.2. Finally, in Section 5.3, we discuss social norms and seek to evaluate social norms and social preferences as possible predictors of SM's reciprocal behavior.

5.1 Relative Importance and Interaction of Intention and Outcome Effects

For each combination of intention and outcome, SM chose by how many points to punish or reward their FM. Negative b's (-3, -2, or -1) represent how many points SM forfeits for

punishing FM by deducting either 3, 6, or 9 of FM's points (i.e., punishment lowers both players' payoffs). Positive b's (2, 4, or 6) represent how many points SM transfers one for one to FM (i.e., a reward represents an efficiency-neutral transfer). Below, we will represent SM's choices by the implied *change* in FM's points (as in Falk *et al.*, 2008) that may be defined as $\Delta(i, j)$, where $i \in \{np, p\}$ and $j \in \{ki, ui\}$. For every possible scenario, $\Delta(\cdot)$ ranges from -9 (strongest punishment) to 6 (highest reward).

5.1.1 The Distribution of Second-Movers' Choices

Figure 2 shows how SMs' choices $\Delta(i, j)$ are distributed.[6] In the case of unkind intentions and the non-preferred outcome (scenario *ui-np*), SMs predominantly chose to reduce the FM's payoff by either 3 or 6 points. In the case of kind intentions and the non-preferred outcome (scenario *ki-np*), SMs chose severe punishments less often and rewards more often. In scenario *ui-p*, compared to scenario *ui-np*, the frequency of punishment is much lower, and many SMs reward the FM. Finally, most SMs rewarded the FM with kind intentions and the preferred outcome (scenario *ki-p*). Still, even in this scenario, more than one-third of SMs punish their FM, predominantly with $\Delta(i, j) = -3$. An intuitive explanation for this pattern is that $\Delta(i, j) = -3$ maximizes SM's monetary payoff and that 72.5% of all SMs who chose $b = -1$ in scenario *ki-p* also adopt it in the other scenarios (i.e., they choose $\Delta(i, j) = -3$ for all i and j) and, hence, play the payoff-maximizing strategy.

5.1.2 The Relative Importance of Intention and Outcome Effects

Figure 3 displays the *average* intention effect (averaged over outcomes *p* and *np*), and the *average* outcome effect (averaged over intentions *ki* and *ui*).[7] FMs earn, on average, 1.34 more points with kind (instead of unkind) intentions. In analogy, FMs earn, on average, 3.19 more points when the preferred (instead of the non-preferred) outcome is drawn. Both effects are significantly different from zero ($p < 0.0001$, Wilcoxon signed-rank test (WSR)). Accordingly, we provide clear evidence of significant intention and outcome effects. The strong change in SM behavior due to FM being kinder and the resulting outcome being more preferred, $\Delta(p, ki) - \Delta(np, ui)$, needs to be attributed to both the change in intentions and the change in the outcome.

Our design allows us to assess how much of the overall effect between the extreme scenarios *ui-np* and *ki-p* can be attributed to kinder intentions and a draw of the preferred outcome. Regarding the relative importance of intentions and outcomes, we observe that the average outcome effect is substantially larger than the average intention effect. The average of $\mathcal{O}(ui)$ and

[6] We pool the data of all treatments. In Section 5.1.4, we will show that (i) all main results hold for each treatment and (ii) treatment differences regarding choice data are absent.

[7] We consider the single effects in Section 5.1.3.

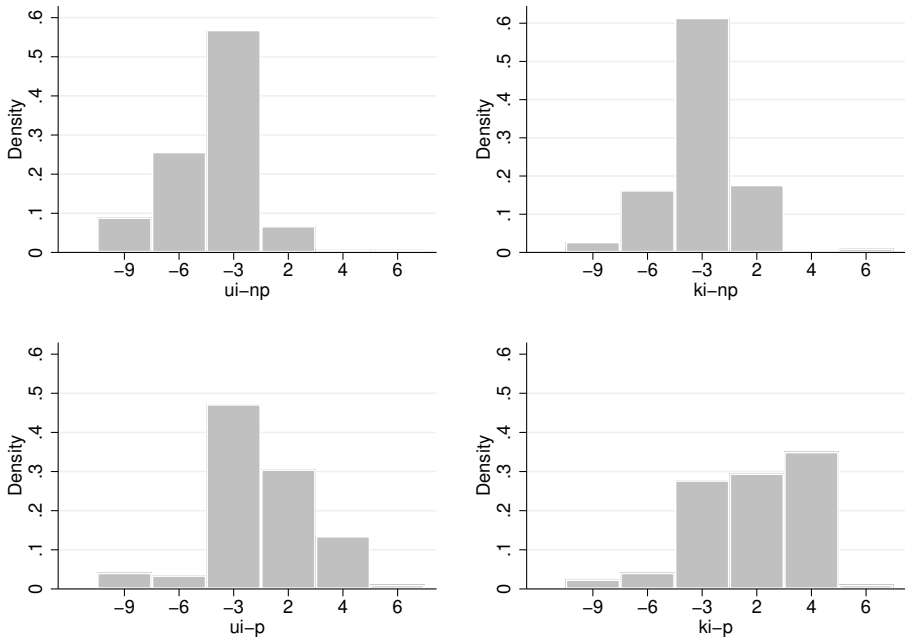


Figure 2: *Distribution of SMs' Choices for Each Scenario (Measured in Terms of the Change of the FM's Payoff)*

$\mathcal{O}(ki)$ is more than twice as large as the average of $\mathcal{I}(p)$ and $\mathcal{I}(np)$, with the difference being highly significant ($p < 0.0001$, WSR). The stronger outcome effect can emerge through a change at the extensive (second-movers being *more likely* to respond at all to a variation in outcomes than a variation in intentions) as well as the intensive margin (second-movers responding *more strongly* to a variation in outcomes than a variation in intentions).

To decompose the effect, Figure 4 shows the SM fraction who react (i) to a variation in outcomes and intentions (i.e., for whom $\mathcal{O}(ki) > 0$ or $\mathcal{O}(ui) > 0$ *and* $\mathcal{I}(p) > 0$ or $\mathcal{I}(np) > 0$), (ii) only to a variation in outcomes (i.e., for whom $\mathcal{I}(p) = \mathcal{I}(np) = 0$ and $\mathcal{O}(ki) > 0$ and/or $\mathcal{O}(ui) > 0$), (iii) only to a variation in intentions (i.e., for whom $\mathcal{O}(ki) = \mathcal{O}(ui) = 0$ and $\mathcal{I}(p) > 0$ and/or $\mathcal{I}(np) > 0$), and (iv) not at all.

The SMs who react to both intention and outcome variations form the largest group (44.79%). Only 23.61% do not adjust their behavior in response to variations in either outcomes or intentions. As argued before, these subjects most likely seek to maximize monetary payoffs: more than 85% of members from this group with no response to an outcome or an intention variation chose the payoff-maximizing strategy: $\Delta(i, j) = -3 \; \forall \; i, j$. The group of subjects who reacted only to outcomes (29.17%) is sizable and much larger than the negligible group of SMs who reacted only to intentions (2.43%). In summary, second-movers are more likely to react to a change in outcomes than to a change in intentions. Among the SMs who react to intentions and outcomes, SMs react more strongly to a variation in outcomes than one in intentions. The average change in the first-mover's payoffs (restricting attention to SMs who react to both outcome

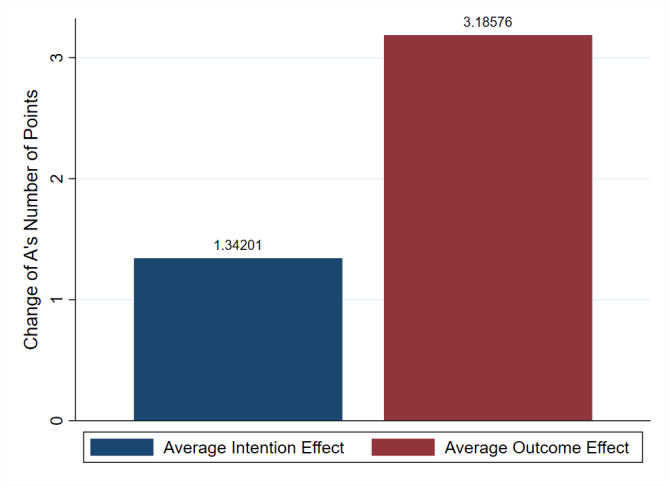


Figure 3: *Average Payoff Effect of Intentions and Outcomes on SM's reactions.*

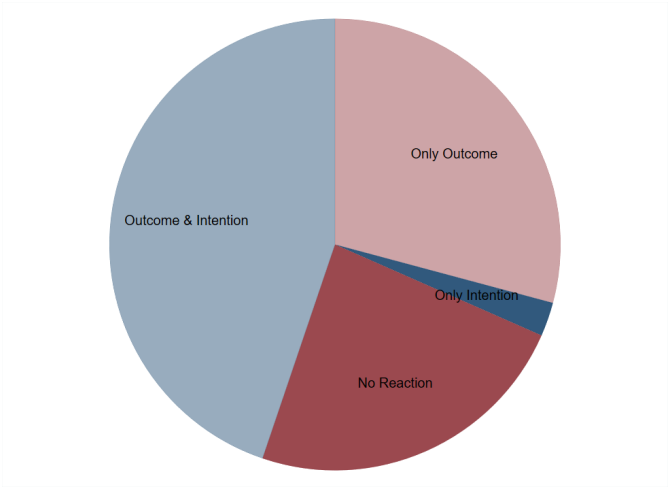


Figure 4: *Distribution of SM Types (Whether They Reacted to Changes in Outcome and/or Intention).*

and intention variations) is significantly larger in response to the preferred outcome than to kind intentions (4.75 vs. 3.95; $p = 0.0066$, WSR).

Result 1. *The average outcome effect is significantly larger than the average intention effect. This is attributable to a stronger change at the extensive and intensive margin.*

Result 1 contrasts with previous results. For example, Falk *et al.* (2008) compare the payoff changes selected by SMs who react to a FM who intentionally and directly implemented (interim) payoffs to the choices of SMs who respond to a randomly drawn outcome. In the intention-free treatment, SMs change interim payoffs much less than SMs who respond to the 13 intention-outcome combinations.[8] The analog of our outcome effect for the case of no in-

[8]In the intention-treatment of Falk *et al.* (2008), every choice of a implements a vector of payoffs that another

tentions (i.e., the change from a FM choice of $a = 2$ to a choice of $a = -2$) amounts to only about 0.09 points.

In Charness and Levine (2007), three (interim) outcomes are possible. The intermediate outcome can be reached by kind and unkind intentions. The best (worst) can exclusively be reached via (un)kind intentions. The authors can determine an intention effect conditional on the intermediate outcome, one outcome effect from a low to an intermediate wage conditional on unkind intentions, and one outcome effect from an intermediate to a high wage conditional on kind intentions. Charness and Levine (2007) conclude that the intention effect is more pronounced in their data.[9]

Why do our results seem to differ from Charness and Levine (2007) and Falk *et al.* (2008)? In our setup, intentions can only shift probability mass between given outcomes, whereas very (un)favorable outcomes can only be reached by (un)kind intentions in Charness and Levine (2007). In fact, the worst possible outcome for the SM conditional on kind intentions is the best possible outcome with unkind intentions. Irrespective of the drawn outcome, this might lead to a more positive evaluation of kind (instead of unkind) intentions and ultimately explain the stronger intention effect.

As compared to Falk *et al.* (2008), in our setup, SM behavior is always a response to (i) an FM choice and (ii) a draw of nature. Hence, we measure the outcome effect in the context of a real interaction, where FM made a payoff-relevant choice. In Falk *et al.* (2008), SM either reacts to FM's direct implementation of an interim outcome (in the intentions treatment) or responds to a random move with player A being a completely passive bystander (in the no-intentions treatment). SMs might perceive these treatments very differently. In the intentions treatment, they might contemplate the optimal reaction for each possible FM choice separately and, hence, engage in the ultimately relevant particular circumstance. In the no-intentions treatment, however, SM may more abstractly consider how to deal with the overall context and realize that counteracting the random draw's impact on the payoff allocation is inefficient: transferring points to FM after a preferred draw is zero-sum whereas deducting points after a non-preferred draw is costly for both players. These aspects might contribute to a lower behavioral variation on the part of SM in the no-intentions treatment.

level of a cannot reach. This leads to 13 intention-outcome combinations. As a result, we cannot create an outcome effect from their data that is conditional on non-neutral intentions.

[9] In the same spirit, Friedrichsen *et al.* (2022) show that principals strongly reward their agent for a high investment (holding constant the project's return) but much less so for the high (instead of low) return (holding the agent's investment constant). However, in this study, the agent's payoffs vary with intentions (due to the investment cost), making it impossible to keep the outcome constant when considering an intention variation.

5.1.3 The Interaction of Intention and Outcome Effects

Having established the relative importance of intentions and outcomes on reciprocal behavior, we now analyze their interaction. In Table 2, we report the results from a linear regression with $\Delta(i, j)$ as the dependent variable and a dummy for kind intentions, the preferred outcome, and the interaction of the two as independent variables. The positive and significant coefficients for kind intentions and the preferred outcome demonstrate a significant intention effect (with the non-preferred outcome as the baseline outcome) and outcome effect (with unkind intentions as the baseline intention). Notably, the positive and significant interaction term demonstrates that a draw of the preferred outcome amplifies the intention effect and that the outcome effect is stronger for kind intentions.

	(1)	(2)
Kind Intentions	1.146***	1.144***
	(0.161)	(0.163)
Preferred Outcome	2.990***	3.007***
	(0.235)	(0.238)
Kind Intentions x Preferred Outcome	0.392**	0.398**
	(0.167)	(0.171)
Constant	-3.774***	-4.247**
	(0.160)	(2.155)
Controls	No	Yes
N	1152	1136
R^2	0.231	0.276

Results from an ordinary least squares regression. The dependent variable is $\Delta(s, i)$.

Standard errors clustered on the subject level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: *Interaction Effect of Outcome and Intentions*

In analogy to Section 5.1.2, the interaction of intentions and outcomes can emerge through changes in behavior at the extensive as well as the intensive margin. First, we find that the share of SMs with a positive outcome effect is significantly higher when intentions were kind instead of unkind (65.6% vs. 56.5%; $p = 0.0001$, McNemar's test). Considering only SMs with a positive outcome effect for both intentions, we find that the strength of the reaction does not depend on intentions being kind or unkind (5.69 vs. 5.88 points, $p = 0.133$, WSR). Second, we find that the frequency of positive intention effects is significantly higher when the preferred outcome materialized (36.5% vs. 28.1%; $p = 0.0009$, McNemar's test). Considering only SMs with a positive intention effect for both outcomes, we find that the reaction's strength does not depend on the draw of the outcome (4.97 vs. 4.89 points, $p = 0.819$, WSR).

Result 2. *We find that $\mathcal{O}(ki) > \mathcal{O}(ui)$ and $\mathcal{I}(p) > \mathcal{I}(np)$. The difference stems from a change in behavior at the extensive margin: more second-movers reward a draw of the preferred outcome conditional on kind intentions.*

Referring back to Figure 2, the result can be illustrated by observing that SMs transfer extensive rewards to their FM only in scenario *ki-p*, i.e., when intentions are kind *and* the preferred outcome applies. An intuitively appealing explanation for this pattern of behavior is motivated reasoning. DeScioli *et al.* (2014) show that peoples' moral and fairness judgments can be influenced by self-interest; that is, people may choose moral standards that are more beneficial for themselves from the set of moral standards that seem acceptable given the circumstances. In our setup, SMs may reason that it is acceptable not to reward their FM as long as at least one dimension, either intention or outcome, is unfavorable to them. This would explain why most second-movers do not reward their FM in scenarios *ui-p* and *ki-np*. In scenario *ki-p*, however, both dimensions turn out to be favorable for SM so that not rewarding FM can no longer be morally justified.

The interaction of intention and outcome effects relates to and challenges existing findings in the outcome-bias literature: while we find that kind intentions amplify the outcome bias, Brownback and Kuhn (2019) find and Gurdal *et al.* (2013) predict that the impact of luck is decreasing with kinder intentions. Friedrichsen *et al.* (2022) do not find outcome biases for either investment success or failure. In a setup in which the second-mover's choice has large efficiency implications, the study by Offerman (2002) produces (i) an intention effect from kind instead of neutral intentions conditional on the good outcome, and (ii) an intention effect resulting from neutral as compared to unkind intentions conditional on the bad outcome. The author emphasizes that effect (ii) is stronger than effect (i) which points in a different direction than our results.

5.1.4 The Impact of a Norm Focus on Intention and Outcome Effects

To assess the robustness of the relative importance and interaction of intention and outcome effects, we conducted treatments aimed at directing SM attention toward the social appropriateness of intentions and/or outcomes. This approach builds on Krupka and Weber (2009) and rests on psychological theories according to which social norms affect behavior only when a person's attention is directed towards the norm.

Figure 5 shows the average $\Delta(i, j)$ for each scenario and treatment separately. All main results from the analysis with pooled data are borne out in each treatment: First, we find an outcome effect ($p < 0.0002$, WSR) and an intention effect ($p < 0.0001$, WSR). Second, the outcome effect dominates the intention effect ($p < 0.01$, WSR). Third, data patterns are consistent with an interaction effect as the outcome effect seems stronger with kind than with unkind intentions; however, it is not statistically significant ($p > 0.122$, WSR).

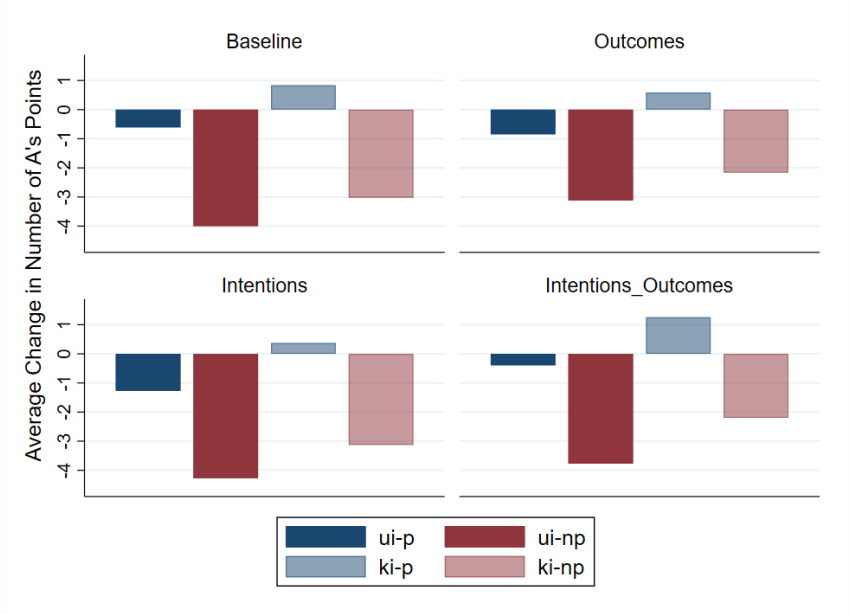


Figure 5: *Average B choices by intentions and outcomes for each treatment*

Importantly, we find no significant treatment effects: although bringing intentions into focus (treatments INT and INTOUT) consistently leads to a slightly larger reaction to kind instead of unkind intentions (1.5 vs. 1.2), the difference is not significant ($p = 0.2574$, WRT). A salient outcome dimension (treatments OUT and INTOUT) does not amplify the impact of outcomes (2.93 vs. 3.44; $p = 0.3829$, WRT). As both differences are statistically insignificant, we conclude that providing a norm focus does not substantially affect the main results.

Result 3. *The presence of outcome and intention effects and both their relative importance and interaction are robust to norm-focus interventions.*

Why are our results robust and similar in size across treatments? Possibly, treatment interventions were effective and shifted attention to social norms but norms were not sufficiently influential in SMs' decision-making. However, the treatment interventions were possibly too weak to direct attention and highlight the respective social appropriateness of kind intentions and the preferred outcome. Although we cannot answer the question conclusively, our post-experimental questionnaire results provide valuable insights. To evaluate the interventions' effectiveness, we asked SMs to rate the social appropriateness of FM's choice options and the interim outcomes on an 11-point Likert scale. Regarding outcomes, the difference in social appropriateness ratings between the preferred and the non-preferred outcome is higher in treatments with an outcome focus, but the difference is insignificant (3.64 vs. 2.68; $p = 0.262$, WRT). The corresponding difference in social appropriateness ratings between kind and unkind intentions is significantly higher in treatments with an intention focus (4.21 vs. 2.89; $p = 0.025$, WRT). Hence, treatments produced a significant shift of social appropriateness ratings in the dimension with the norm focus and the direction. This suggests that the influence of norms on

SMs' choices may be small, a subject we turn to in section 5.3.

5.2 First-Mover's Choice of Probability Distribution

We find that 60.2% of FMs chose unkind intentions, with female subjects more likely to do so than male subjects (64% vs. 53.4%; $p = 0.005$, Fisher's exact).[10] FM's choice of unkind intentions renders the non-preferred outcome more likely, and thus appears selfish. However, choosing unkind intentions did not pay off for FM. The average final payoff of a FM implementing unkind intentions is significantly lower than that of a FM choosing kind intentions ($11.49€$ vs. $11.93€$; $p = 0.0021$, Wilcoxon rank-sum test (WRT)). Several factors contribute to the unattractiveness of choosing unkind intentions for first-movers. First, via the intention effect, FMs are more likely to get punished. Second, the non-preferred outcome is much more likely after the choice of unkind intentions which, again, increases the chance of punishment. Finally, due to the interaction effect, FMs are more likely to receive a reward after the draw of the preferred outcome mainly when intentions are kind.

5.3 Predicting the Second-Mover's Choice Using Social Norms and Social Preferences

This section examines the drivers behind SM's choices, focusing on assessing the role played by social norms and social preferences. We will begin by describing and analyzing the social norms elicited from third parties. Subsequently, we will link the norms data with the choice data to evaluate the explanatory power of social norms compared to social preferences.

5.3.1 Social Norms Regarding Second-Mover's Choice

For each scenario, participants rated the appropriateness of the six SM actions, which we denote $r(b|i,j)$. Possible responses range from "very socially inappropriate" (coded as $r = -1$) to "very socially appropriate" (coded as $r = 1$). Figure 6 illustrates the mean appropriateness ratings of SM's actions, separated by scenario. Notably, regardless of the scenario, a reward emerges as the most appropriate SM action. Even when unkind intentions and the non-preferred outcome apply, punishing the FM is not deemed most appropriate. This contributes evidence to the research question on the existence of a social norm regarding punishment raised by Fehr and Schurtenberger (2018). Despite frequent punishment by SMs, there is no discernible social norm advocating punishment.

[10] The results from a logistic regression with kind intentions as the dependent variable confirm the gender difference when controlling for personal characteristics and traits (see Table A.1 in the Appendix). Additionally, we observe that FM's choice of kind intentions is positively related to measures of trust, norm obedience, and prosociality, while it is negatively correlated with age.

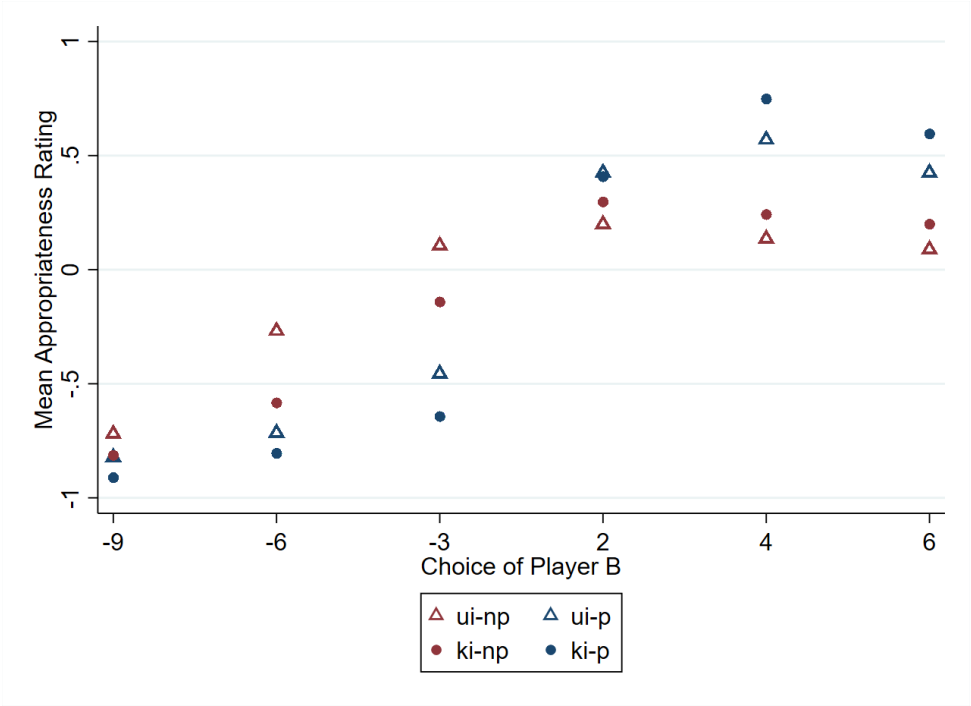


Figure 6: *Mean Appropriateness of Second-Mover Actions by Scenario*

Aggregating the punishment ratings

$$\mathcal{P}(i,j) = r(-9|i,j) + r(-3|i,j) + r(-1|i,j)$$

and the reward ratings

$$\mathcal{R}(i,j) = r(2|i,j) + r(4|i,j) + r(6|i,j)$$

and comparing $\mathcal{P}(i,j)$ and $\mathcal{R}(i,j)$, we find that a draw of the preferred as compared to the non-preferred outcome (comparing blue and red symbols in Figure 6) renders rewards more and punishments less appropriate for both intentions ($p < 0.0001$ for every comparison, WSR).[11] This provides evidence of an outcome-dependent social norm for SM behavior.[12] Regarding kind versus unkind intentions (comparing triangles and circles of identical color in Figure 6), a significant negative effect on the appropriateness of punishments ($p < 0.042$) is observed for both outcomes, but the positive effect on the appropriateness of rewards is insignificant ($p > 0.415$). Thus, kind intentions reduce the appropriateness of FM punishment but do not positively affect the appropriateness of rewards. This provides evidence of a partially intentions-dependent social norm of SM behavior.

For a given scenario (i,j), the difference between the aggregated social appropriateness of reward and the aggregated social appropriateness of punishment choices, $\mathcal{R}(i,j) - \mathcal{P}(i,j)$,

[11] The corresponding graphical representation can be found in Figure A.1 in the appendix.

[12] For example, Gächter *et al.* (2017) also consider summary measures from their norms data when they consider an action (in)appropriate when the majority ranks the action either "very (in)appropriate", "somewhat (in)appropriate", or "(in)appropriate".

indicates how much more appropriate rewards are compared to punishments. These differences allow us to derive the *norms-outcome* and the *norms-intention* effects. The results reported above indicate that the norms-outcome effect emerges via significantly more appropriate reward *and* significantly less appropriate punishment choices, while the norm-intention effect is shaped solely by less appropriate punishment choices. In other words, outcomes affect the social norm more than intentions. This is confirmed by a statistical test on the equality of aggregated norms-outcome and norm-intention effects ($p < 0.0001$, WSR).[13]

In our data, the norms-intention and the norms-outcome effects do not interact. Unlike our choice data, the norms-outcome effect does not depend on Player A's intentions. Equivalently, the norms intention effect is not affected by the draw of the outcome ($p = 0.4677$, WSR).

Result 4. *Social norms foresee an adjustment of SM behavior in response to either an outcome variation for fixed intentions or an intention variation for a fixed outcome, where the former adjustment exceeds the latter. The norms-outcome and the norms-intention effect do not interact.*

5.3.2 Second-Mover's Choices: The Predictive Power of Social Norms and Social Preferences

To assess the explanatory power of social norms for SMs' behavior and compare it to that of social preferences, we employ a conditional (fixed-effects) logistic regression, following the approach of Krupka and Weber (2013) and Gächter *et al.* (2013). The dependent variable is a dummy that, for each action (in a given scenario), takes value one if SM selected it and zero otherwise. The probability of selecting an action is assumed to depend on the utility associated with that action relative to the utility associated with all other actions:

$$\Pr(b = k) = \frac{\exp\{U(k)\}}{\sum_{\Delta=-9,\ldots,6} \exp\{U(\Delta)\}}; \;\; k = -9, \ldots, 6.$$

As a benchmark, we consider a model in which SMs care (only) about their own monetary payoff, imposing a linear restriction on the utility of money such that

$$U(b) = \beta_1 \Pi_i^B(b).$$

Next, we augment the model with a preference for norm compliance. For every b, we include the mean social appropriateness rating $N(b)$ from the norms experiment. Hence,

$$U(b) = \beta_1 \Pi_i^B(b) + \beta_2 N(b).$$

[13] Note that this difference is not only born out in the aggregated appropriateness ratings but also shows up in the individual choices options: by and large, the differences between equal-color ratings are smaller than the corresponding differences between equal-marker ratings.

To compare the explanatory power of social norms to that of social preferences, we adopt the model of Charness and Rabin (2002) and, hence, allow SM to have outcome-based and intentions-based social preferences. The utility function reads

$$U(b) = \beta_1 \Pi_i^B(b) + \beta_3 \min\{\Pi_i^A(b) - \Pi_i^B(b), 0\} + \beta_4 \max\{\Pi_i^A(b) - \Pi_i^B(b), 0\} + \beta_5 \mathbb{1}_{\text{Urn L}} \left(\Pi_i^B(b) - \Pi_i^A(b)\right).$$

The coefficients β_3 (β_4) capture the impact of outcome-based preferences via (dis)advantageous payoff inequity. Intentions-based reciprocity is captured by the coefficient β_5. Charness and Rabin (2002) assume that, next to distributional preferences, the utility increases with the payoff difference $\Pi_i^B - \Pi_i^A$ if the other player misbehaved. We assume FM misbehaves if she shows unkind intentions.

	(1)	(2)	(3)	(4)	(5)
Own Payoff (β_1)	0.0727***	0.0811***	0.0863***	0.0869***	0.240***
	(0.00285)	(0.00268)	(0.00756)	(0.00759)	(0.0421)
Appropriateness Rating (β_2)		0.396***			-0.978***
		(0.0655)			(0.274)
Advantageous Payoff Difference (β_3)			0.125***	0.152***	0.293***
			(0.00931)	(0.0113)	(0.0359)
Disadvantageous Payoff Difference (β_4)			-0.165***	-0.144***	-0.0701***
			(0.0132)	(0.0131)	(0.0228)
Reciprocity (β_5)				0.0479***	0.0595***
				(0.00575)	(0.00550)
N	6912	6912	6912	6912	6912
(Pseudo) R^2	0.018	0.026	0.098	0.102	0.105

Notes: Results from conditional logistic regressions. The dependent variable is a dummy indicating the chosen SM action. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: *Determinants of Reciprocal Behavior*

The result in Column (1) of Table 3 documents that actions with a higher Π_i^B are more likely to be chosen. The results in Column (2) suggest that SMs follow social norms: SMs are more likely to choose socially appropriate (compared to less appropriate) actions. Thus, a preference to comply with social norms can explain part of the observed SM behavior. In Column (3), we restrict attention to inequity aversion. The positive and significant coefficient β_3 indicates that SMs seek to reduce advantageous payoff inequity ($\pi_i^A - \pi_i^B < 0$), while the negative and significant coefficient β_4 indicates a preference for lower disadvantageous payoff inequity ($\pi_i^A - \pi_i^B > 0$). In Column (4), we additionally account for reciprocity and find a positive and significant coefficient β_5. An unkind FM strengthens competitive preferences of SM.

Considering the Pseudo R^2 from Columns (2) and (4), we find that the social preference model explains the data better than the model incorporating social norm ratings. Our first result on the relative explanatory power thus reads as follows:

Result 5. *The incorporation of social preferences improves the explanatory power of the empirical model more than the incorporation of social norms.*

We employ two approaches to further compare the explanatory power of social norms and social preferences. First, we present results from a model encompassing both social norms and social preferences. Subsequently, we identify situations where predictions derived from a preference for norm compliance and social preferences diverge. We then analyze the actual behavior of second-movers.

In Column (5) of Table 3, the results from a model incorporating social norms and social preferences are displayed. Notably, the social preference coefficients maintain their sign and significance, whereas the coefficient of social appropriateness becomes *significant and negative*.

To explain the negative coefficient of social appropriateness, we delve into the determinants of the social appropriateness rating at the individual level. Table 4 presents findings from ordinary least squares regression in Column (1) with the standardized norm rating as the dependent variable, and a conditional logit model in Column (2), where the dependent variable is a dummy equal to one if the participant assigned the highest appropriateness rating to the action in question and zero otherwise.

Examining an *advantageous* payoff difference, reducing it by rewarding FMs is deemed socially appropriate: the positive coefficient $\hat{\beta}_2$ indicates that, when being ahead, choices that lead to higher FM payoffs are more appropriate for a given SM payoff. This normative rating aligns with SMs' behavior. Conversely, reducing a *disadvantageous* payoff difference by punishing FMs is not deemed socially appropriate, as indicated by the positive coefficient $\hat{\beta}_3$: even when being behind, the social appropriateness of SM's actions is increasing in FM's payoff. This normative rating conflicts with SMs' behavior. Overall, while SMs are motivated by inequity aversion, the social norm aligns with social efficiency: for advantageous as well as disadvantageous payoff differences, redistribution via rewards is considered more socially appropriate than costly, inefficient punishment. We can use our findings from Table 4 to return to the question about a social norm of punishment:

Result 6. *Social norms reflect social efficiency concerns and, thus, exclude a social norm of punishment in our setup.*

This finding suggests that the procedure for eliciting social norms is valid, yet there are limits to its explanatory power. The asymmetry in how disadvantageous payoff differences influence the social appropriateness of second-mover behavior, in contrast to observed behavior, contributes to explaining the negative coefficient of social appropriateness in Column (5) of Table 3.

	(1) Norm Rating	(2) Max Rating
Own Payoff ($\hat{\beta}_1$)	0.148***	0.334***
	(0.00731)	(0.0446)
Advantageous Payoff Difference ($\hat{\beta}_2$)	0.133***	0.385***
	(0.00538)	(0.0468)
Disadvantageous Payoff Difference ($\hat{\beta}_3$)	0.0728***	0.193***
	(0.00690)	(0.0307)
Reciprocity ($\hat{\beta}_4$)	0.00897***	0.0269*
	(0.00303)	(0.0137)
Constant	-1.522***	
	(0.0834)	
N	2256	2232
(Pseudo) R^2	0.463	0.237

Results from an ordinary least squares regression in Column (1). The dependent variable is the standardized norm rating. Results from a conditional logistic regression in Column (2). The dependent variable is a dummy variable for the action with the highest appropriateness rating. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: *Determinants of the Social Appropriateness Rating of Different Reciprocal Actions.*

This can be best illustrated in scenarios with the non-preferred outcome, where second-movers are confronted with disadvantageous payoff inequity. As social efficiency calls for rewards instead of punishment, the social appropriateness of rewards exceeds the corresponding ratings of punishment choices. In line with SMs' inequity aversion, however, punishment choices are more likely to be chosen than rewards. Hence, social-norm considerations predict the opposite behavior of what is observed. In summary, our comparisons of the predictive power of social norms and social preferences favor the latter, aligning with the results from Section 5.1.4, indicating that changes in social appropriateness ratings after norm focus interventions do not alter actual SM choices.

6 Conclusion

Reciprocity plays a crucial role in human behavior, with individuals reciprocating others' actions based on perceived outcomes and intentions. This study provides experimental evidence on the relative importance and interaction of intention and outcome effects in reciprocal behavior. In contrast to previous studies, our research disentangles these two determinants and

presents their relative importance and interaction.

We establish that the impact of outcomes on reciprocal behavior dominates that of intentions. This is remarkable as first-movers are primarily punished or rewarded based on a lottery draw instead of their choice. Our second main result is that the influence of the outcome on the reciprocal behavior depends on the first-mover's intention and vice versa. The outcome effect is more pronounced when intentions are kind instead of unkind.

The relative importance and interaction of outcomes and intentions in determining reciprocal behavior are relevant in various domains. For instance, in responding to a customer complaint, should a firm offer monetary compensation (focusing on outcomes) or explain precautions against product failures (focusing on intentions) to avoid retaliatory action? Similarly, after a policy failure, should politicians try to convince voters of their benevolent intentions or offer concessions in other policy domains for voter compensation to secure re-election prospects? Our results suggest that focusing on outcomes might be more successful in inducing a favorable reaction. Moreover, our findings cast doubt on whether an organization's ex-ante fair promotion process (good intention) can secure morale and be evaluated as fair ex-post by eligible but non-promoted employees (e.g., Baron and Kreps, 1999).

Examining what drives second-movers' reciprocal behavior, we consider social norms and social preferences. While social norms reflect the relevance of intention and outcome effects, their explanatory power falls short of that of social preferences. This asymmetry can be understood by considering how second-movers *ought* to respond to disadvantageous inequity according to social norms and how they *do* respond. Social norms point away from punishment due to social efficiency concerns, but second-movers tend to punish in circumstances of disadvantageous inequity, consistent with inequity aversion. This implies that there is no social norm of punishment, as punishment is not more socially appropriate than rewards, even when the first-mover shows unkind intentions *and* the disadvantageous outcome applies.

APPENDIX

MATHEMATICAL DERIVATIONS

We first derive the optimal SM choice $\Delta^*(i,j)$ for each scenario (i,j). To determine the parameter region in which a specific level of Δ is optimal, the utility associated with that level needs to be compared to adjacent levels. If SM has no incentive to deviate one step upwards (downwards) from Δ, she has no incentive to deviate in that direction. Assuming that SM chooses the larger Δ when indifferent, pairwise comparisons reveal that the optimal choices after kind intentions, $\Delta^*(p,ki)$ and $\Delta^*(np,ki)$, are given by:

$$\Delta^*(p,ki) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} \\ -3 & \text{if } -\frac{1}{2} \leq \rho < \frac{1}{6} \\ 2 & \text{if } \frac{1}{6} \leq \rho < \frac{1}{2} \\ 4 & \text{if } \sigma < \frac{1}{2} \leq \rho \\ 6 & \text{if } \sigma \geq \frac{1}{2} \end{cases} \quad \Delta^*(np,ki) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} \\ -6 & \text{if } \sigma < -\frac{1}{2} \leq \rho \\ -3 & \text{if } -\frac{1}{2} \leq \sigma < \frac{1}{6} \\ 2 & \text{if } \frac{1}{6} \leq \sigma < \frac{1}{2} \\ 6 & \text{if } \sigma \geq \frac{1}{2} \end{cases}$$

Likewise, optimal choices can be derived after unkind intentions, $\Delta^*(p,ui)$ and $\Delta^*(np,ui)$. Clearly, the level of the parameter θ is important in this context. We find that the effect of θ is similar to a translation of the conditions regarding the parameters ρ and σ. For $\theta \leq \frac{1}{2}$, we get:

$$\Delta^*(p,ui) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} + \theta \\ -3 & \text{if } -\frac{1}{2} + \theta \leq \rho < \frac{1}{6} + \theta \\ 2 & \text{if } \frac{1}{6} + \theta \leq \rho < \frac{1}{2} + \theta \\ 4 & \text{if } \sigma < \frac{1}{2} + \theta \leq \rho \\ 6 & \text{if } \sigma \geq \frac{1}{2} + \theta \end{cases} \quad \Delta^*(np,ui) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} + \theta \\ -6 & \text{if } \sigma < -\frac{1}{2} + \theta \leq \rho \\ -3 & \text{if } -\frac{1}{2} + \theta \leq \sigma < \frac{1}{6} + \theta \\ 2 & \text{if } \frac{1}{6} + \theta \leq \sigma < \frac{1}{2} + \theta \\ 6 & \text{if } \sigma \geq \frac{1}{2} + \theta \end{cases}$$

If $\theta \in \left(\frac{1}{2}, \frac{5}{6}\right]$, SM avoids any transfers larger than 2. Hence:

$$\Delta^*(p,ui) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} + \theta \\ -3 & \text{if } -\frac{1}{2} + \theta \leq \rho < \frac{1}{6} + \theta \\ 2 & \text{if } \rho \geq \frac{1}{6} + \theta \end{cases} \quad \Delta^*(np,ui) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} + \theta \\ -6 & \text{if } \sigma < -\frac{1}{2} + \theta \leq \rho \\ -3 & \text{if } -\frac{1}{2} + \theta \leq \sigma < \frac{1}{6} + \theta \\ 2 & \text{if } \sigma \geq \frac{1}{6} + \theta \end{cases}$$

Finally, if $\theta > \frac{5}{6}$, SM chooses only negative levels of Δ. It follows that

$$\Delta^*(p,ui) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} + \theta \\ -3 & \text{if } \rho \geq -\frac{1}{2} + \theta \end{cases} \quad \Delta^*(np,ui) = \begin{cases} -9 & \text{if } \rho < -\frac{1}{2} + \theta \\ -6 & \text{if } \sigma < -\frac{1}{2} + \theta \leq \rho \\ -3 & \text{if } \sigma \geq -\frac{1}{2} + \theta \end{cases}$$

INTENTION EFFECT:

Regarding the intention effect, note that the impact of unkind intentions on SM's utility due to $\theta(\Pi_p^{SM} - \Pi_p^{SM})$, which consists of decreasing the attractiveness of $\Delta > \hat{\Delta}$ and increasing the attractiveness of $\Delta < \hat{\Delta}$, is decreasing in Δ. In other words, a scenario with unkind instead of kind intentions makes raising punishment and decreasing transfers more attractive. If SM changes behavior as a reaction to unkind intentions, she will move to lower Δ.

The intention effect, given a draw of the preferred outcome,

$$\mathcal{I}(p) = \Delta^*(p,ki) - \Delta^*(p,ui)$$

depends on SM's (i) outcome-based preferences (as measured by ρ and σ where $\rho \geq \sigma$) and (ii) intention-based preferences (as measured by θ). In the following, we consider the different ranges relevant for $\Delta^*(p,ki)$ and then consider the implied intention effects for the various levels of θ.

- $\rho < -\frac{1}{2}$: We find $\Delta^*(p,ki) = \Delta^*(p,ui) = -9$ so $\mathcal{I}(p) = 0$.
- $-\frac{1}{2} \leq \rho < \frac{1}{6}$: We have $\Delta^*(p,ki) = -3$ and

$$\mathcal{I}(p) = \begin{cases} 0 & \text{if } \theta \leq \rho + \frac{1}{2} \\ 6 & \text{if } \theta > \rho + \frac{1}{2} \end{cases}.$$

- $\frac{1}{6} \leq \rho < \frac{1}{2}$: Using $\Delta^*(p,ki) = 2$, we get

$$\mathcal{I}(p) = \begin{cases} 0 & \text{if } \theta \leq \rho - \frac{1}{6} \\ 5 & \text{if } \rho - \frac{1}{6} < \theta \leq \rho + \frac{1}{2} \\ 11 & \text{if } \theta > \rho + \frac{1}{2} \end{cases}.$$

- $\sigma < \frac{1}{2} \leq \rho$: Using $\Delta^*(p,ki) = 4$, we get

$$\mathcal{I}(p) = \begin{cases} 0 & \text{if } \theta \leq \rho - \frac{1}{2} \\ 2 & \text{if } \rho - \frac{1}{2} < \theta \leq \rho - \frac{1}{6} \\ 7 & \text{if } \theta > \rho - \frac{1}{6} \end{cases}.$$

- $\sigma \geq \frac{1}{2}$: With $\Delta^*(p, ki) = 6$, it follows that

$$\mathcal{I}(p) = \begin{cases} 0 & \text{if } \theta \leq \sigma - \frac{1}{2} \\ 2 & \text{if } \sigma - \frac{1}{2} < \theta \leq \rho - \frac{1}{2} \\ 4 & \text{if } \rho - \frac{1}{2} < \theta \leq \rho - \frac{1}{6} \\ 9 & \text{if } \theta > \rho - \frac{1}{6} \end{cases}.$$

Next, we turn to the intention effect, given a draw of the non-preferred outcome,

$$\mathcal{I}(np) = \Delta^*(np, ki) - \Delta^*(np, ui).$$

It depends on SM's outcome- and intention-based preferences as follows:

- $\rho < -\frac{1}{2}$: $\Delta^*(np, ki) = \Delta^*(np, ui) = -9$ so that $\mathcal{I}(np) = 0$.
- $\sigma < -\frac{1}{2} \leq \rho$: With $\Delta^*(np, ki) = -6$, it follows that

$$\mathcal{I}(np) = \begin{cases} 0 & \text{if } \theta \leq \rho + \frac{1}{2} \\ 3 & \text{if } \theta > \rho + \frac{1}{2} \end{cases}.$$

- $-\frac{1}{2} \leq \sigma < \frac{1}{6}$: Using $\Delta^*(np, ki) = -3$, we get

$$\mathcal{I}(np) = \begin{cases} 0 & \text{if } \theta \leq \sigma + \frac{1}{2} \\ 3 & \text{if } \sigma + \frac{1}{2} < \theta \leq \rho + \frac{1}{2} \\ 6 & \text{if } \theta > \rho + \frac{1}{2} \end{cases}.$$

- $\frac{1}{6} \leq \sigma < \frac{1}{2}$: Using $\Delta^*(np, ki) = 4$, it follows that

$$\mathcal{I}(np) = \begin{cases} 0 & \text{if } \theta \leq \sigma - \frac{1}{6} \\ 5 & \text{if } \sigma - \frac{1}{6} < \theta \leq \sigma + \frac{1}{2} \\ 8 & \text{if } \sigma + \frac{1}{2} < \theta \leq \rho + \frac{1}{2} \\ 11 & \text{if } \theta > \rho + \frac{1}{2} \end{cases}.$$

- $\sigma \geq \frac{1}{2}$: With $\Delta^*(np, ki) = 6$, it results that

$$\mathcal{I}(np) = \begin{cases} 0 & \text{if } \theta \leq \sigma - \frac{1}{2} \\ 4 & \text{if } \sigma - \frac{1}{2} < \theta \leq \sigma - \frac{1}{6} \\ 9 & \text{if } \theta > \sigma - \frac{1}{6} \end{cases}.$$

The above shows that the intention effects $\mathcal{I}(p)$ and $\mathcal{I}(np)$ are strictly positive when $\rho > -\frac{1}{2}$ and θ sufficiently high.

OUTCOME EFFECT:

The outcome effect conditional on kind intentions is defined as

$$\mathcal{O}(ki) = \Delta^*(p, ki) - \Delta^*(np, ki).$$

It depends on the parameters regarding advantageous and disadvantageous inequity in the following way:

$$\mathcal{O}(ki) = \begin{cases} 0 & \text{if } \rho < -\frac{1}{2} \\ 3 & \text{if } -\frac{1}{2} \leq \rho < \frac{1}{6} \text{ and } \sigma < -\frac{1}{2} \\ 8 & \text{if } \frac{1}{6} \leq \rho < \frac{1}{2} \text{ and } \sigma < -\frac{1}{2} \\ 10 & \text{if } \rho \geq \frac{1}{2} \text{ and } \sigma < -\frac{1}{2} \\ 0 & \text{if } \rho < \frac{1}{6} \text{ and } \sigma \geq -\frac{1}{2} \\ 5 & \text{if } \frac{1}{6} \leq \rho < \frac{1}{2} \text{ and } -\frac{1}{2} \leq \sigma < \frac{1}{6} \\ 7 & \text{if } \rho \geq \frac{1}{2} \text{ and } -\frac{1}{2} \leq \sigma < \frac{1}{6} \\ 0 & \text{if } \rho < \frac{1}{2} \text{ and } \sigma \geq \frac{1}{6} \\ 2 & \text{if } \rho \geq \frac{1}{2} \text{ and } \frac{1}{6} \leq \sigma < \frac{1}{2} \\ 0 & \text{if } \sigma \geq \frac{1}{2} \end{cases}$$

There exist SM types that do not change their Δ when the preferred outcome is drawn instead of the non-preferred one. The remaining types react by choosing a higher Δ. Accordingly, the expected outcome effect is positive.

Next, we derive the outcome effect for unkind intentions. We consider the intervals for θ used when specifying $\Delta^*(p, ui)$ and $\Delta^*(np, ui)$ above in turn:

- $\theta \leq \frac{1}{2}$: We get

$$\mathcal{O}_I(ui) = \begin{cases} 0 & \text{if } \rho < -\frac{1}{2} + \theta \\ 3 & \text{if } -\frac{1}{2} + \theta \leq \rho < \frac{1}{6} + \theta \text{ and } \sigma < -\frac{1}{2} + \theta \\ 8 & \text{if } \frac{1}{6} + \theta \leq \rho < \frac{1}{2} + \theta \text{ and } \sigma < -\frac{1}{2} + \theta \\ 10 & \text{if } \rho \geq \frac{1}{2} + \theta \text{ and } \sigma < -\frac{1}{2} + \theta \\ 0 & \text{if } \rho < \frac{1}{6} + \theta \text{ and } \sigma \geq -\frac{1}{2} + \theta \\ 5 & \text{if } \frac{1}{6} + \theta \leq \rho < \frac{1}{2} + \theta \text{ and } -\frac{1}{2} + \theta \leq \sigma < \frac{1}{6} + \theta \\ 7 & \text{if } \rho \geq \frac{1}{2} + \theta \text{ and } -\frac{1}{2} + \theta \leq \sigma < \frac{1}{6} + \theta \\ 0 & \text{if } \rho < \frac{1}{2} + \theta \text{ and } \sigma \geq \frac{1}{6} + \theta \\ 2 & \text{if } \rho \geq \frac{1}{2} + \theta \text{ and } \frac{1}{6} + \theta \leq \sigma < \frac{1}{2} + \theta \\ 0 & \text{if } \sigma \geq \frac{1}{2} + \theta \end{cases}$$

- $\theta \in \left(\frac{1}{2}, \frac{5}{6}\right]$: The outcome effect amounts to

$$\mathcal{O}_{II}(ui) = \begin{cases} 0 & \text{if } \rho < -\frac{1}{2} + \theta \\ 3 & \text{if } -\frac{1}{2} + \theta \leq \rho < \frac{1}{6} + \theta \text{ and } \sigma < -\frac{1}{2} + \theta \\ 8 & \text{if } \rho \geq \frac{1}{6} + \theta \text{ and } \sigma < -\frac{1}{2} + \theta \\ 0 & \text{if } \rho < \frac{1}{6} + \theta \text{ and } \sigma \geq -\frac{1}{2} + \theta \\ 5 & \text{if } \rho \geq \frac{1}{6} + \theta \text{ and } -\frac{1}{2} + \theta \leq \sigma < \frac{1}{6} + \theta \\ 0 & \text{if } \sigma \geq \frac{1}{6} + \theta \end{cases}$$

- $\theta > \frac{5}{6}$: It follows that

$$\mathcal{O}_{III}(ui) = \begin{cases} 0 & \text{if } \rho < -\frac{1}{2} + \theta \\ 3 & \text{if } \sigma < -\frac{1}{2} + \theta \leq \rho \\ 0 & \text{if } \sigma \geq -\frac{1}{2} + \theta \end{cases}$$

Independent of θ, the outcome effect for unkind intentions, $\mathcal{O}(ui)$, is weakly positive.

INTERACTION EFFECT:

Regarding the interaction of intentions and outcomes in shaping punishment choices, we aim to compare $\mathcal{O}(ki)$ to $\mathcal{O}(ui)$. For a given SM type, the impact of unkind intentions on the outcome effect, as measured via the difference

$$\mathcal{M} = \mathcal{O}(ki) - \mathcal{O}(ui)$$

can be positive or negative. As an example, first consider a type with $\rho \in \left(-\frac{1}{2}+\theta, \frac{1}{6}\right)$ and $\sigma \in \left(-\frac{1}{2}, -\frac{1}{2}+\theta\right)$: as $\Delta^*(np, ki) = \Delta^*(p, ki) = -3$ and $\Delta^*(np, ui) = -6 < -3 = \Delta^*(p, ui)$, SM reacts to the non-preferred outcome by increasing punishment only after FM has shown unkind intentions. Second, consider $\rho \in \left(-\frac{1}{2}, -\frac{1}{2}+\theta\right)$ and $\sigma < -\frac{1}{2}$: with $\Delta^*(np, ki) = -6 < -3 = \Delta^*(p, ki)$ and $\Delta^*(np, ui) = \Delta^*(p, ui) = -9$, SM reacts to the non-preferred outcome by increasing punishment only after FM has shown kind intentions. While unkind intentions strengthen the outcome effect of the first type, they mitigate it for the second. Accordingly, the sign of the interaction is type-specific.

Although proving a general interaction effect valid for every SM type is impossible, we will adopt three approaches to derive clear-cut predictions. First, we will show that, under a uniform distribution of ρ-σ-types, the expected outcome effect under kind intentions exceeds that under unkind intentions. Formally, we will show that, for any given $\theta > 0$, $E\left[\mathcal{O}(ki)\right] > E\left[\mathcal{O}(ui)\right]$. This approach uses the outcome effect sizes for the different SM types, as measured by $\Delta^*(p, j) - \Delta^*(np, j)$. To show that the result does not hinge on effect sizes, we adopt a second approach by analyzing a dummy variable equal to one when an SM type shows a positive outcome effect. Comparing the parameter region in which SMs react to the variation in the interim outcome in scenarios with kind intentions to the region resulting under unkind intentions shows that more SM types show a positive outcome effect after kind than after unkind intentions. Finally, we will show that, for θ sufficiently large, the interaction is consistently (weakly) positive for all SM types with inequity aversion or competitive preferences.

First, assume a uniform distribution of SM types over possible ρ-σ combinations. The ex-ante likelihood that a SM will show a specific outcome effect is determined by the area of preference parameters ρ and σ that induces this outcome effect relative to the total area of all parameter combinations (which is 2). As an example, all players with $\rho \in \left[-\frac{1}{2}, \frac{1}{6}\right)$ and $\sigma \in \left[-1, -\frac{1}{2}\right)$ have $\mathcal{O}(ki) = 3$. Accordingly, the probability attached to this outcome effect is

$$\frac{\left[\frac{1}{6} - \left(-\frac{1}{2}\right)\right]\left[-\frac{1}{2} - (-1)\right]}{2} = \frac{1}{2}\frac{2}{3}\frac{1}{2} = \frac{1}{6}.$$

The expected outcome effect under kind intentions amounts to

$$\mathbb{E}\left[\mathcal{O}(ki)\right] = \frac{1}{2}\left[\frac{2}{3}\frac{1}{2}3 + \frac{1}{2}\frac{1}{3}8 + \frac{1}{2}\frac{1}{2}10 + \frac{2}{3}\frac{1}{3}5 + \frac{2}{3}\frac{1}{2}7 + \frac{1}{3}\frac{1}{2}2\right] = \frac{155}{36}.$$

The expected outcome effect under unkind intentions can be written as

$$\mathbb{E}\left[\mathcal{O}(ui)\right] = \begin{cases} \mathbb{E}\left[\mathcal{O}_I(ui)\right] & \text{if } \theta \leq \frac{1}{2} \\ \mathbb{E}\left[\mathcal{O}_{II}(ui)\right] & \text{if } \frac{1}{2} < \theta \leq \frac{5}{6} \\ \mathbb{E}\left[\mathcal{O}_{III}(ui)\right] & \text{if } \theta > \frac{5}{6} \end{cases},$$

where

$$\begin{aligned}\mathbb{E}\left[\mathcal{O}_I(ui)\right] =& \frac{1}{2}\left[\left(\frac{1}{2}+\theta\right)\frac{2}{3}3+\left(\frac{1}{2}+\theta\right)\frac{1}{3}8+\left(\frac{1}{2}+\theta\right)\left(\frac{1}{2}-\theta\right)10\right.\\ &\left.+\frac{2}{3}\frac{1}{3}5+\frac{2}{3}\left(\frac{1}{2}-\theta\right)7+\frac{1}{3}\left(\frac{1}{2}-\theta\right)2\right]\\ =& \frac{155}{36}-\frac{1}{3}\theta-5\theta^2,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathcal{O}_{II}(ui)\right] =& \frac{1}{2}\left[\left(\frac{1}{2}+\theta\right)\frac{2}{3}3+\left(\frac{1}{2}+\theta\right)\left(\frac{5}{6}-\theta\right)8+\frac{2}{3}\left(\frac{5}{6}-\theta\right)5\right]\\ =& \frac{32}{9}+\frac{2}{3}\theta-4\theta^2,\end{aligned}$$

and

$$\mathbb{E}\left[\mathcal{O}_{III}(ui)\right] = \frac{1}{2}\left[\left(\frac{1}{2}+\theta\right)\left(\frac{3}{2}-\theta\right)3\right] = \frac{9}{8}+\frac{3}{2}\theta-\frac{3}{2}\theta^2.$$

As $\frac{\partial\mathbb{E}[\mathcal{O}(ui)]}{\partial\theta} < 0$ and $\mathbb{E}\left[\mathcal{O}(ui)\right]\big|_{\theta=0} = \mathbb{E}\left[\mathcal{O}(ki)\right]$, it follows that $\mathbb{E}\left[\mathcal{O}(ui)\right] < \mathbb{E}\left[\mathcal{O}(ki)\right] \forall\theta > 0$.

Second, we use a binary variable that takes the value 1 for an SM type if and only if the outcome effect for this type is strictly positive, i.e., $\Delta^*(np,j)-\Delta^*(p,j) > 0$. We apply the same analysis as in the first approach, replacing the outcome effect sizes with the value of the binary variable. This corresponds to comparing the share of SM types experiencing a strictly positive outcome effect with kind intentions to that experiencing it with unkind intentions. Denoting this share by $\mathcal{S}\left[\mathcal{O}(j)\right]$, we obtain

$$\mathcal{S}\left[\mathcal{O}(ki)\right] = \frac{53}{72}$$

for the scenario with kind intentions. Accordingly, roughly 74% of SM types exhibit an outcome effect after kind intentions. With unkind intentions, the share amounts to

$$\mathcal{S}\left[\mathcal{O}(ui)\right] = \begin{cases}\frac{53}{72}-\frac{1}{2}\theta^2 & \text{if } \theta \leq \frac{1}{2}\\ \frac{47}{72}+\frac{1}{6}\theta-\frac{1}{2}\theta^2 & \text{if } \frac{1}{2} < \theta \leq \frac{5}{6}\\ \frac{3}{8}+\frac{1}{2}\theta\left(1-\theta\right) & \text{if } \theta > \frac{5}{6}\end{cases}.$$

Again, as $\frac{\partial\mathcal{S}[\mathcal{O}(ui)]}{\partial\theta} < 0$ and $\mathcal{S}\left[\mathcal{O}(ui)\right]\big|_{\theta=0} = \mathcal{S}\left[\mathcal{O}(ki)\right]$, it follows that $\mathcal{S}\left[\mathcal{O}(ui)\right] < \mathcal{S}\left[\mathcal{O}(ki)\right] \forall\theta > 0$.

Third, restrict attention to all competitive SM types, for whom $\sigma \leq \rho < 0$. We have shown that $\mathcal{O}(ui) = 0$ if $\rho < -\frac{1}{2}+\theta$. If $\theta > \frac{1}{2}$, this holds for all competitive types. As $\mathcal{O}(ki) > 0$ if

$\sigma < -\frac{1}{2} \leq \rho$, it holds that the outcome effect is strictly positive for some SM types. Hence, the interaction effect is (weakly) positive for all competitive SM types if $\theta > \frac{1}{2}$. Next, consider all inequity-averse SM types. For these types, $\sigma < 0 < \rho$ holds. Assume $\theta > \frac{5}{6}$. The outcome effect under unkind intentions is then strictly positive for some SM types. In particular, $\mathcal{O}(ui) = 3$ if $\rho \geq -\frac{1}{2} + \theta$. With $\theta > \frac{5}{6}$, this requires $\rho \geq \frac{1}{3}$, which in turn ensures $\mathcal{O}(ki) > 3$. Hence, the interaction effect is (weakly) positive for all inequity-averse SM types if $\theta > \frac{5}{6}$.

ADDITIONAL FIGURES AND TABLES

	FM Showing Kind Intentions	
	(1)	(2)
Male	0.453***	0.585***
	(0.136)	(0.146)
Age		-0.0161*
		(0.00888)
Risk		0.0348
		(0.0334)
Trust		0.0567*
		(0.0291)
Norm Obedience		0.0501*
		(0.0280)
SVO		0.451**
		(0.176)
Constant	-0.435***	-3.213**
	(0.0858)	(1.279)
N	369	364
pseudo R^2	0.022	0.063

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.1: *The Determinants of Kind Intentions*

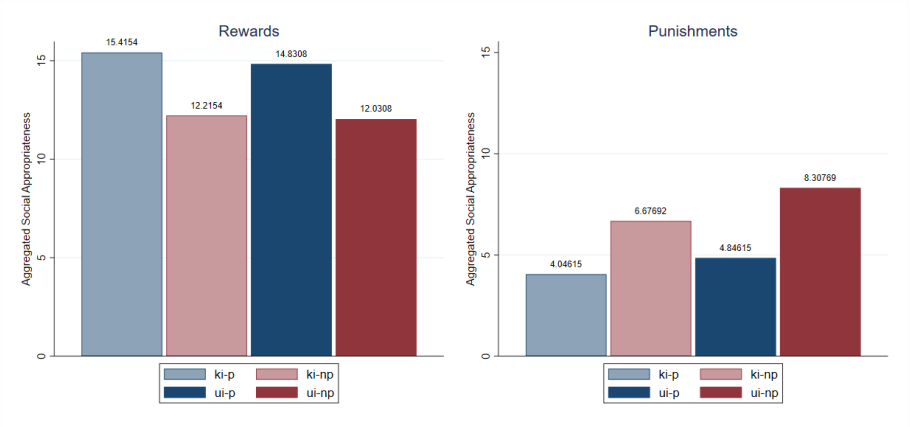


Figure A.1: *Aggregated social appropriateness ratings of reward (left panel) and punishment (right panel) choices by intentions and outcomes.*

INSTRUCTIONS

INSTRUCTIONS CHOICE EXPERIMENT

Welcome to the experiment

Today, you will take part in a study on decision-making behavior. This experiment consists of an interactive part and a questionnaire. For your participation and answering the questionnaire, you will receive 4€. In addition, you can earn extra money in the experiment. The amount you earn depends on your decisions and those of other people.
You have 10 minutes to read and edit each page. The time remaining for the current page is displayed in the yellow box. You cannot continue participating in the experiment if you exceed the time.

Please click Next to continue.

Overview of the experiment:

The structure of the experiment is described below. This description gives you an overview of the experiment and does not give all the details yet. On the following pages, the experiment will be described in detail and you will get all necessary information.

In this experiment, there are two roles, A and B. The roles are assigned randomly. During the experiment, you will be assigned to another real person. Each pair consists of a real person in role A and a real person in role B. Both persons have an identical initial endowment of points and make decisions that affect the points of both persons.

In Stage 1, A decides whether a ball is drawn from urn L or urn R. Both urns contain red and black balls but have different compositions. The color of the drawn ball decides how the payoffs of both persons change at the end of Stage 1. If a black ball is drawn, points are deducted from A, and B is assigned three times those points. In contrast, when a red ball is drawn, A is assigned additional points, while precisely the same number of points are deducted from B. The probability of a red ball is much higher for urn L.

B can directly influence the payoffs of both persons in stage 2, conditioning on the urn chosen by A and the relevant ball color. B can give up points to have precisely the same number credited to A. Alternatively, B can invest his points to have three times the number of points invested by B deducted from A.

Please click Next to continue.

Role:

On the following pages you will find a detailed description of the different roles and the procedure of the experiment. There are two roles in this experiment, A and B. The roles are assigned randomly. During the experiment you will be assigned to another real existing person. Each pair consists of a real person in the role of A and a real person in the role of B. Both persons have an identical initial endowment of 12 points.

You have the A/B role.

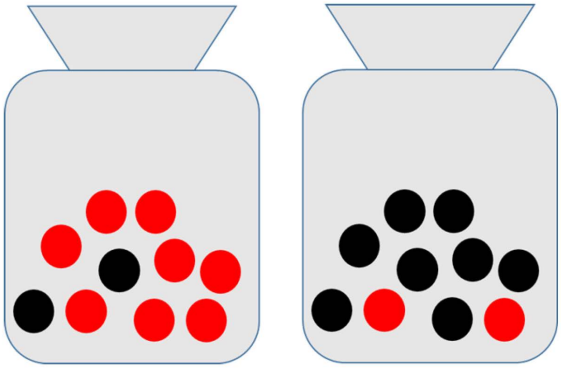
Neither before nor after the experiment will you learn the identity of the person assigned to you. The person assigned to you also learns nothing about your identity. The final score, which is relevant for your payout, is influenced by your decisions and by the decisions of the person assigned to you in the role B/A. The exact process can be divided into two stages.

Please click Next to continue.

Stage 1:

A decides from which of two urns, L or R, a ball will be drawn. The urns contain black and red balls. The color of the drawn ball determines how the scores of both players change at the end of Stage 1.

- A black ball is drawn: A is deducted 2 points and B is assigned 6 points.



- A red ball is drawn: A is assigned 2 points and B is deducted 2 points.

There are ten balls in both urns, but different numbers of red and black balls. In urn L, there are 8 red and 2 black balls. In urn R, there are 2 red and 8 black balls. In other words, the probability that a red ball is drawn from urn L is 80%. In contrast, the probability that a red ball is drawn from urn R is only 20%. Correspondingly, the probability that a black ball is drawn from urn L is 20%, and the probability that a black ball is drawn from urn R is 80%. The urns L and R are shown in the following figure.

The scores of both persons at the end of Stage 1 thus depend on whether a red or a black ball is drawn. After A has determined the urn, the computer draws a ball according to the probabilities of the selected urn.

Please click Next to continue.

Stage 2:

In stage 2, B can directly change both persons' payoffs, reacting to the urn chosen by A and the ball color drawn.

- B can give up points to transfer them to A: B can give up 2, 4, or 6 points to A, which increases A's score by 2, 4 or 6 points.
- B can invest own points to reduce A's score: For each point invested, A's score is reduced by three points. B can invest 1, 2, or 3 points to reduce A's score by 3, 6, or 9 points.

B's six alternatives are summarized in the table below, with the resulting score changes for both individuals.

Please click Next to continue.

B's Action Set	-3	-2	-1	2	4	6
Change in score B	-3	-2	-1	-2	-4	-6
Change in score A	-9	-6	-3	+2	+4	+6

Stage 2:

B decides for each scenario defined by the urn choice of A and the drawn ball color (i.e., the payoffs at the end of Stage 1). The choice is made using the following table:

Scenario	B's Action Set					
	-3	-2	-1	2	4	6
A chose **urn L** (8 red & 2 black balls), and one **black ball** was drawn; that is, A was deducted 2 points, and B was assigned 6 points	o	o	o	o	o	o
A chose **urn R** (2 red & 8 black balls), and one **black ball** was drawn; that is, A was deducted 2 points, and B was assigned 6 points	o	o	o	o	o	o
A chose **urn R** (2 red & 8 black balls), and a **red ball** was drawn; that is, A was assigned 2 points, and B was deducted 2 points	o	o	o	o	o	o
A chose **urn L** (8 red & 2 black balls), and a **red ball** was drawn; that is, A was assigned 2 points, and B was deducted 2 points	o	o	o	o	o	o

Please click Next to continue.

Payoff:

After A has chosen either urn L or urn R, a ball is drawn from this urn according to the associated probabilities. In stage 2, B has made a score change decision for each possible scenario at the end of Stage 1; that is, for each possible combination of A's urn choice and the color of the drawn ball. The decision made by B for the actual relevant scenario (that is, the urn chosen by A and the ball drawn from it) is now implemented.

Before making the own decision, B is not informed about the urn chosen by A and the color of the ball drawn. Because B decides for every possible scenario and in the end only the decision made for the actual relevant scenario is implemented, it corresponds to the situation in which B is informed about A's choice of urn and the drawn ball before making his own decision. Accordingly, in Stage 2, B reacts to the urn chosen by A and the color of the drawn ball.

The final score results from the drawn ball's color and B's decision for the relevant combination of A's urn choice and ball color. Your final payout from this decision situation results from your final score, where 1 point corresponds to a payout of 0.75€. You will also receive the safe amount of 4€ for showing up on time and answering the questionnaire at the end of the experiment.

Please click Next to continue.

For Type A:

Your decision:

After answering the control questions, we now ask you to make the decision relevant to your payout. Choose between urn L and urn R:

Urn L	Urn R
o	o

Please click Next to continue.

For Type B:

Treatment INT and INTOUT

Assessment Urn Choice:

Before deciding on score changes in stage 2, please read the following text carefully and answer the question below. Note that the person assigned to you in role A will not see this text and question.

If a black ball is drawn, the total score, i.e., the sum of the points of A and B, increases by four

points at the end of Stage 1. In contrast, when a red ball is drawn, the total number of points remains unchanged. If A chooses urn R, the probability of increasing the total score is 80%.

However, if A chooses urn L, the probability of increasing the total score is much lower, only 20%.

What do you think is the socially desired urn choice of A?

Urn L	o
Urn R	o

Please click Next to continue.

For Type B:

Treatment OUT and INTOUT

Assessment outcome Stage 1:

Now we ask you to read the following text carefully and answer the question below. Note that the person assigned to you in role A will not see this text and question.

If a black ball is drawn, this increases the total score, i.e., the sum of the points of A and B, by four points at the end of Stage 1.

In contrast, if a red ball is drawn, the total score remains unchanged at the end of Stage 1.

What do you think is the socially desired outcome of the urn choice?

Black ball	o
Red ball	o

Please click Next to continue.

For Type B:

Your decision:

After answering the control questions, we ask you to make the decisions relevant to your payout. Select the desired score change for each combination of urn choice of A and ball color.

Your decision if	**-3**	**-2**	**-1**	**2**	**4**	**6**
A chose **urn L,** and a **black ball** was drawn; that is, A was deducted 2 points, and B was assigned 6 points	o	o	o	o	o	o
A chose **urn R,** and a **black ball** was drawn; that is, A was deducted 2 points, and B was assigned 6 points	o	o	o	o	o	o
A chose **urn R,** and a **red ball** was drawn; that is, A was assigned 2 points, and B was deducted 2 points	o	o	o	o	o	o
A chose **urn L** and a **red ball** was drawn; that is, A was assigned 2 points, and B was deducted 2 points	o	o	o	o	o	o

INSTRUCTIONS NORMS EXPERIMENT

Welcome to the experiment:

You are taking part in a study on decision-making behavior. For your participation, you will receive 4€. In addition, you can earn extra money in the experiment. The amount you earn depends on your decisions and those of other people.

Please click Next to continue.

Overview of the experiment:

In this experiment, you will evaluate the actions of other persons. First, the situation in which the persons choose a particular action is explained to you. Next, you will be shown the alternative actions the persons choose from.

For each alternative action a person might choose, your task is to decide whether, in that situation, the action is socially appropriate and consistent with moral and appropriate social behavior or socially inappropriate and inconsistent with moral and appropriate social behavior. By socially appropriate, we mean behavior that most people consider correct and ethical. To get an

idea of the task, we first present an example.

Before we describe the relevant situation in detail, we will illustrate your task in a simple example. You will not make any assessment in this example. The example only serves to explain the task to you.

Please click Next to continue.

Example situation:

Suppose person X is sitting in a café near campus and notices that a guest has forgotten his wallet. Person X can now choose from four alternative actions:

1. Take the wallet with you.
2. Address other guests.
3. Leave the wallet behind.
4. Hand over the wallet to the café's owner.

Below you can see the table with the alternative actions. You would select one of the six possible ratings for each action of very socially inappropriate, socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, socially appropriate, or very socially appropriate by checking a box.

	Very socially inappropriate	Socially inappropriate	Somewhat socially inappropriate	Somewhat socially appropriate	Socially appropriate	Very socially appropriate
Take the wallet with you						
Address other guests						
Leave the wallet behind						
Hand over the wallet to the café's owner						

Please click Next to continue.

Example situation:

If you feel that taking the wallet with you is very socially inappropriate, approaching other patrons is socially appropriate, leaving the wallet behind is somewhat socially inappropriate, and handing the wallet to the café's owner is very socially appropriate, then you would enter your opinion into the table as follows:

	Very socially inappropriate	Socially inappropriate	Somewhat socially inappropriate	Somewhat socially appropriate	Socially appropriate	Very socially appropriate
Take the wallet with you	X					
Address other guests					X	
Leave the wallet behind			X			
Hand over the wallet to the café's owner						X

Please click Next to continue.

Introduction

After the example, you will now be informed about the situation relevant to you and the alternatives to be evaluated by you. Your task is judging the alternative actions of participants in a real experiment according to their social appropriateness or inappropriateness. To make this assessment, you will read the same description the participants receive during the experiment.

After reading about the situation and alternatives, you will learn how your payoff depends on your and the other participants' choices before you submit your evaluations.

Please click Next to continue.

Participants are shown the exact instructions as in the Choice experiment (except for norm focus interventions).

Your task: Assessment of the alternative courses of action

You have now been shown the same description participants receive in the experiment. Your task is to decide for the situation at hand whether an alternative action is socially appropriate and consistent with moral and appropriate social behavior or socially inappropriate and inconsistent with moral and appropriate social behavior.

You first assess the alternative actions for participants in role A. Then you assess the alternative actions for participants in role B. These are assessed for each scenario, that is, for each combination of A's urn choice and the color of the drawn ball.

Please click Next to continue.

Your payout

After you have submitted your assessments, the computer randomly selects an action. The evaluations made for this action by all session participants determine your payoff. If your assessment of the social appropriateness of this action matches the assessment most frequently given by the other people in the session, you will receive an additional payout of 20€. Otherwise, you will not receive an additional payout.

Example: The computer randomly selects the action For the situation where A selects urn L (8 red and 2 black balls) and a red ball is drawn (A is assigned 2 points and B is deducted 2 points), B selects -3 (A is assigned 9 and B is deducted 3 points). If your assessment would have been somewhat socially inappropriate, then you would earn an additional 20€ if somewhat socially inappropriate was the most frequently mentioned assessment by the other session participants. Otherwise, you will not receive an additional payout.

Please click Next to continue.

Your assessment: Participant in role A

Evaluate how socially appropriate you rate participant A's actions:

- A chooses urn L, that is, with 80% probability, 2 points are assigned to A and 2 points are deducted from B; with 20% probability, 2 points are subtracted from A and 6 points are assigned to B.
- A chooses urn R, that is, with 20% probability, 2 points are assigned to A and 2 points are deducted from B; with 80% probability, 2 points are subtracted from A and 6 points

are assigned to B.

	Very socially inappropriate	Socially inappropriate	Somewhat socially inappropriate	Somewhat socially appropriate	Socially appropriate	Very socially appropriate
A selects urn L						
A selects urn R						

Please click Next to continue.

Your assessment: Participant in role B (scenario 1)

Rate how socially appropriate you find the participants' alternative actions in role B in the following situation:

A chose urn L (8 red & 2 black balls), and a red ball was drawn; that is, A was assigned 2 points and B was deducted 2 points.

Alternative Actions of B	**-3**	**-2**	**-1**	**2**	**4**	**6**
Change in score B	-3	-2	-1	-2	-4	-6
Change in score A	-9	-6	-3	+2	+4	+6

Please click Next to continue.

Your assessment: Participant in role B (scenario 2)

Rate how socially appropriate you find the participants' alternative actions in role B in the following situation:

A chose urn L (8 red & 2 black balls), and a black ball was drawn; that is, A was deducted 2 points and B was assigned 6 points.

(same Table as for scenario 1)

	Very socially inappropriate	Socially inappropriate	Somewhat socially inappropriate	Somewhat socially appropriate	Socially appropriate	Very socially appropriate
B selects -3.						
B selects -2						
B selects -1						
B selects 2						
B selects 4						
B selects 6						

Please click Next to continue

Your assessment: Participants in role B (scenario 3)

Rate how socially appropriate you find the participants' alternative actions in role B in the following situation:

A chose urn R (2 red & 8 black balls), and a red ball was drawn; that is, A was assigned 2 points, and B was deducted 2 points.

(same Table as for scenario 1)

Please click Next to continue.

Your assessment: Participant in role B (scenario 4)

Rate how socially appropriate you find the participants' alternative actions in role B in the following situation:

A chose urn R (2 red & 8 black balls), and a black ball was drawn; that is, A was deducted 2 points, and B was assigned 6 points.

(same Table as for scenario 1)

REFERENCES

ABBINK, K., IRLENBUSCH, B. and RENNER, E. (2000). The moonlighting game – an experimental study on reciprocity and retribution. *Journal of Economic Behavior and Organization*, **42**, 265–77.

BARON, J. and HERSHEY, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, **54**, 569–79.

BARON, J. N. and KREPS, D. M. (1999). *Strategic Human Resources: Frameworks for General Managers*. London: Wiley.

BIZER, G., MAGIN, R. and LEVINE, M. (2014). The social-norm espousal scale. *Personality and Individual Differences*, **58**, 106 – 11.

BOCK, O., BAETGE, I. and NICKLISCH, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, **71**, 117–120.

BROWNBACK, A. and KUHN, M. E. (2019). Understanding outcome bias. *Games and Economic Behavior*, **117**, 342–60.

CHAN, N. and WOLK, L. (2023). Reciprocity with stochastic loss. *Journal of the Economic Science Association*, **9**, 51–65.

CHANG, D., CHEN, R. and KRUPKA, E. (2019). Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior*, **116**, 158–178.

CHARNESS, G. (2004). Attribution and reciprocity in a simulated labor market: An experimental investigation. *Journal of Labour Economics*, **22**, 665–88.

— and LEVINE, D. I. (2007). Intention and stochastic outcomes: An experimental study. *The Economic Journal*, **117** (522), 1051–1072.

— and RABIN, M. (2002). Understanding social preferences with simple tests. *The quarterly journal of economics*, **117** (3), 817–869.

CHEN, D. L., SCHONGER, M. and WICKENS, C. (2016). otree – an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, **9**, 88–97.

COHN, A., FEHR, E. and GOETTE, L. (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, **61** (8), 1777–1794.

COX, J. C., SADIRAJ, K. and SADIRAJ, V. (2008). Implications of trust, fear, and reciprocity for modeling economic behavior. *Experimental Economics*, **11**, 1–24.

CROCKETT, M., ÖZDEMIR, Y. and FEHR, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, **143**, 2279–86.

DAVIS, B. J., KERSCHBAMER, R. and OEXL, R. (2017). Is reciprocity really outcome-based? a second look at gift-exchange with random shocks. *Journal of the Economic Science Association*, **3**, 149–60.

DESCIOLI, P., MASSENKOFF, M., SHAW, A., PETERSEN, M. B. and KURZBAN, R. (2014). Equity or equality? moral judgments follow the money. *Proceedings of the Royal Society B: Biological Sciences*, **281** (1797), 20142112.

FALK, A., BECKER, A., DOHMEN, T., ENKE, B., HUFFMAN, D. and SUNDE, U. (2018). Global evidence on economic preferences. *Quarterly Journal of Economics*, **133**, 1645–92.

—, FEHR, E. and FISCHBACHER, U. (2003). On the nature of fair behavior. *Economic Inquiry*, **41**, 20–26.

—, — and — (2008). Testing theories of fairness – intentions matter. *Games and Economic Behavior*, **62** (1), 287–303.

FEHR, E. and SCHMIDT, K. (1999). A theory of fairness, competition and co-operation. *Quarterly Journal of Economics*, **114**, 817–68.

— and SCHURTENBERGER, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, **2** (7), 458–468.

FRIEDRICHSEN, J., MOMSEN, K. and PIASENTI, S. (2022). Ignorance, intention and stochastic outcomes. *Journal of Behavioral and Experimental Economics*, **100**, 101913.

FRIEHE, T. and UTIKAL, V. (2018). Intentions under cover–hiding intentions is considered unfair. *Journal of Behavioral and Experimental Economics*, **73**, 11–21.

GURDAL, M. Y., MILLER, J. B. and RUSTICHINI, A. (2013). Why blame? *Journal of Political Economy*, **121** (6), 1205–1247.

GÄCHTER, S., GERHARDS, L. and NOSENZO, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, **97**, 72–86.

—, NOSENZO, D. and SEFTON, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences? *Journal of the European Economic Association*, **11**, 548–73.

KERSCHBAMER, R. and OEXL, R. (2023). The effect of random shocks on reciprocal behavior in dynamic principal-agent settings. *Experimental Economics*, **26**, 468 – 88.

KIMBROUGH, E. O. and VOSTROKNUTOV, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, **14**, 608–38.

KRUPKA, E. and WEBER, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, **30** (3), 307–320.

KRUPKA, E. L. and WEBER, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, **11** (3), 495–524.

MCCABE, K. A., RIGDON, M. L. and SMITH, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, **52**, 267–75.

MURPHY, R. O., ACKERMANN, K. A. and HANDGRAAF, M. (2011). Measuring social value orientation. *Judgment and Decision making*, **6** (8), 771–781.

OFFERMAN, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, **46** (8), 1423–1437.

RAND, D. G., FUDENBERG, D. and DREBER, A. (2015). It's the thought that counts: The role of intentions in noisy repeated games. *Journal of Economic Behavior and Organization*, **116**, 481–99.

RUBIN, J. and SHEREMETA, R. (2016). Principal-agent settings with random shocks. *Games and Economic Behavior*, **117**, 342–60.

TOUSSAERT, S. (2017). Intention-based reciprocity and signaling of intentions. *Journal of Economic Behavior and Organization*, **137**, 132–44.

TRAXLER, C. and WINTER, J. (2012). Survey evidence on conditional norm enforcement. *European Journal of Political Economy*, **28**, 390 – 98.

VAN DER WEELE, J. J., KULISA, J., KOSFELD, M. and FRIEBEL, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American economic Journal: microeconomics*, **6** (3), 256–264.

XIAO, E. and KUNREUTHER, H. (2016). Punishment and cooperation in stochastic social dilemmas. *Journal of Conflict Resolution*, **60** (4), 670–93.