Freie Universität Berlin

FUELS WORKING PAPER #13

# Compensation and Incentives to Breach

**Sven Hoeppner**
Charles University, Law School, Department for Economics; CERGE-EI Foundation; Free University Empirical Legal Studies Center (FUELS) and University of Zurich, Department of Sociology

**Freie Universität Empirical Legal Studies Center (FUELS)**
Freie Universität Berlin
Fachbereich Rechtswissenschaft
Department of Law
http://fuels.berlin

# Compensation and Incentives to Breach

**ABSTRACT**
I conduct a controlled experiment that provides an institution test on how compensatory damages affect participants' willingness to forego a more lucrative transaction. Efficient breach theory recommends expectation damages as a remedy for breach of contract. Prior literature shows that people tend to honor their agreements and see transactions through at private costs. I elicit the reserve price of second movers to not play cooperatively with first movers in a trust game and instead engage in another transaction. The results show that contractual remedies reduce agents' reserve price, i.e., agents are more likely to forego the transaction with the principal under compensation and other remedies than without. Moreover, under compensatory damages, principals anticipate this effect. Finally, the compensation mechanism decreases principals' punishment of opportunistic agents. Compensatory damages thus bolster the behavioral foundations of efficient breach theory, provide institutional debiasing, and assure contract remedies' indemnification and vindication function.

KEYWORDS
breach of contract, compensation, willingness-to-breach, agreement, institutional debiasing, experimental economics, Bayesian statistics

JEL CLASSIFICATIONS
C11, C91, D02, D86, D90, K12

**Sven Hoeppner**
Assistant Professor
Charles University, Law School, Department for Economics
nám. Curieových 901/7
116 40 Prague, Czech Republic
e-mail: hoeppnes@prf.cuni.cz

FUELS
EMPIRICAL LEGAL STUDIES CENTER

# Compensation and Incentives to Breach

Sven Hoeppner*

Working Paper
(Version: 16 April 2024)

* Sven Hoeppner is Assistant Professor at Charles University, Law School, Department for Economics, nám. Curieových 901/7, 116 40 Prague, Czech Republic. Email: hoeppnes@prf.cuni.cz. Sven Hoeppner is CERGE-EI Foundation Teaching Fellow and Senior Research Fellow at Free University Empirical Legal Studies Center (FUELS) and at the University of Zurich, Department of Sociology.

I am grateful to Fred Bereskin, Ben Depoorter, Andreas Engert, Eric Helland, Svenja Hippel, Wlada Mill, Heinrich Nax, Holger Spamann, Alexander Stremitzer, and Eric Talley for providing excellent comments and critical feedback, both of which improved the paper tremendously. Discussion with participants at the Annual Conference on Empirical Legal Studies (CELS) in Toronto, the Annual Conference of the German Law and Economics Association (GLEA) in Berlin, the Annual Conference of the German Association for Experimental Economics Research (GFEW) in Magdeburg, and the Law in Action Seminar Series at Charles University in Prague further helped to refine the paper. Finally, I am indebted to Christiana Christodoulou for invaluable technical assistance.

I use R version 4.0.3 (2020-10-10) for statistical analysis (R Core Team, 2018). Central parts of the statistical analysis rely on *JAGS 4.3.0* (Plummer, 2003) and *runjags 2.0.4-6* (Denwood, 2016). The data and the full code is available upon request from the author and included in the supplementary materials on the paper's website at: [Insert link to paper website here.].

# 1 Introduction

Economic relationships are often governed, formally or informally, by contracts. Contracts allow parties to commit to a particular course of action, and given credible commitment, parties can agree on mutually beneficial transactions. However, as the relationship evolves and future uncertainty unravels, changing circumstances and new outside options may render the prior commitment relatively invaluable. In these cases, society could be better off if tied-up resources could easily be committed to new alternative uses.

Efficient breach to the rescue. According to the simple efficient-breach theory, promisors can opt to renege from an obligation and pay expectation damages to the promisee if doing so would lead to more productive use of their resources and, thus, improve social welfare. Accordingly, under (perfectly compensatory) expectation damages, contract parties have (Pareto) efficient incentives to breach.[1]

The theory of efficient breach rests on a subtle behavioral assumption: harnessing the efficiency gains of contract breach assumes that contract parties disregard agreed-upon obligations and act opportunistically once a more profitable course of action becomes available. Contrary to this standard economics assumption, mounting empirical evidence suggests that people are much more cooperative than classic economic theory holds. In contractual settings, particularly, experimental evidence suggests a behavioral force of promise-keeping. Even if statements of intent are mere cheap talk, such promissory commitment enhances subsequent levels of trust and cooperation (Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006, 2010, 2011; Vanberg, 2008; Sutter, 2009). In a recent paper, Ederer and Stremitzer (2017) disentangle the two leading explanations for the behavioral effects of promises. One explanation is that promises move the promisor's beliefs about the promisee's expectations. Guilt-averse promisors suffer disutility from disappointing others' expectations. As a result, they keep their promise to avoid guilt (Charness and Dufwenberg, 2006; Ederer and Stremitzer, 2017). The other explanation is that promisors have a belief-independent preference for promise-keeping per se (Vanberg, 2008; Sutter, 2009; Charness and Dufwenberg, 2010), which can be modeled with a psychological cost of being inconsistent (e.g.: Ellingsen and Johannesson, 2004) or a cost of lying (e.g.: Chen et al., 2008; Kartik, 2009).[2] The specific explanation notwithstanding, if contract parties incur psychological costs when breaking a promise, opportunities for efficient breach may remain unexploited. For instance, Wilkinson-Ryan (2015) elicits participants' willingness-to-breach an agreement. Her results show that participants in her experiment require a substantial premium to break prior agreements for profit. In other words, contrary to the efficient-breach theory, they do not disregard prior commitments for any small profit.

Efficient breach theory predicates a remedy, most notably expectation damages (cf.: Klass, 2014; Cooter and Ulen, 2016; Eisenberg, 2018).[3] However, prior commitment as a concern for

[1] While the efficient-breach theory is a central concept in the economic theory of contracts and contract law, it spurred much discussion. The refined theory is much more nuanced regarding the costs and benefits of the decision to breach (cf.: Klass, 2014; Eisenberg, 2018).

[2] This explanation also links promissory commitment to topics of consistent behavior in social psychology (cf.: Festinger, 1957; Cialdini, 2007).

[3] Of course, the efficient result can obtain under specific performance (see only: Ulen, 1984). Specific performance will trigger a bargaining process in which the promisor and the promisee divide the surplus of efficient breach of the agreement. The efficient outcome under specific performance, however, requires that the conditions

efficient breach has not been studied with contractual remedies present. While Wilkinson-Ryan (2015) finds that promisors require a substantial premium to breach for profit, promisors in her experiment make their breach decision absent remedies. By contrast, the broader experimental literature on the effects of contract remedies does not focus on the effect of remedies on the propensity to breach. Bigoni et al. (2017) conducted an incentivized experiment to investigate the effects of specific performance on the compensation required by the promisee to accept a gain-seeking breach or a loss-avoiding breach. Similarly, while the experiment of Depoorter and Tontrup (2012) allows for efficient breach, the authors focus on enforcing the original contract by the promisee. Finally, while Hoeppner et al. (2017) investigate opportunistic behavior in a trust game under different remedies, the authors do not elicit the second mover's willingness-to-breach given a profitable outside opportunity but focus on the moral hazard problem endogenous to the relationship between principal and agent.

This article reports the results of an incentivized experiment employed as an institution test to study the effect of expectation damages on agents' willingness-to-breach (WTB). Sampling 510 participants, I study how different remedies influence an agent's reserve price to renege on a prior commitment. I compare a decision environment without consequences for non-cooperative play with an environment featuring perfectly compensatory expectation damages. As compensation implies a transfer from the promisor to the promisee, I add two treatments with fines and recovery mechanisms to disentangle sources of potential effects. While the experiment is an institution test, I obtain a richer data set, including information about possible behavioral channels, by providing principals with an opportunity to punish agents for their choice indirectly and by looking at post-experimental survey responses.

The experiment yields exciting and vital results. On the aggregate level, agents display an increased WTB when a breach triggers compensation of the principal or when a breach results in a fine. By contrast, I find no effects when principals recover their losses upon breach. To a somewhat lesser degree, I observe the same pattern for aggregate principals' beliefs about the agents' WTB, i.e., principals under a compensation mechanism or a fine have elevated beliefs about the agents' WTB compared to principals in the recovery regime or absent a remedy. Interestingly, compensatory damages substantially reduce the gap between principals' beliefs and agents' actual WTB. On the aggregate level, principals' expectations about agents' behavior align much better with actual agents' behavior under compensatory damages than any other remedy. The experiment uncovers corresponding results on the individual level. Independent of whether the agent perceives the interaction with the principal as being based on an agreement, compensations and fines reduce the agents' reserve prices for breach. By contrast, the recovery mechanism's effect depends on the agent's assessment. Like the other remedies, the recovery procedure decreases the agent's reserve price absent perceived agreement with the principal. However, if the agent perceives the interaction as founded on an agreement, the recovery mechanism increases the agent's reserve price for deciding opportunistically. Finally, absent perceived agreement, principals under compensation or recovery mechanisms administer less punishment for opportunistic behavior. Given perceived agreement, however, principals under the exact mechanisms punish agents more harshly for breach.

for Coasean bargaining hold and that the difference between the subjective value that the promisee places on performance and the outside option of the promisor leave a bargaining space.

These findings suggest that under expected damages and fines, agents' behavior is much more in line with the assumptions of efficient breach theory. Moreover, as compensation elevates principals' beliefs about the agents' WTB, the difference diminishes between both. The alignment between principals' beliefs and agents' choices implies that compensation mechanisms also create a form of institutional debiasing. The results also provide evidence for the indemnification and vindication function of compensatory damages. As compensation and recovery mechanisms reduce the punishment rate for opportunistic choices, principals evaluate breach as less morally reprehensible. Finally, the results imply that contract remedies and agreements are substitutes when principals' commitment decisions are concerned.

The paper is structured as follows. The next Section 2 describes the experimental design. Section 3 places this study in the context of related literature and derives hypotheses. Section 4 presents the results. Finally, section 5 discusses the findings before Section 6 concludes.

# 2 Experiment Design & Procedures

## 2.1 Baseline Condition & Game Features

The experiment aims to create opportunities to break an agreement for profit to elicit participants' willingness-to-breach (WTB), i.e., the monetary amount required to renege on a previous commitment. I employ a minimal version of the trust game to facilitate such ex-post opportunism.

This version of the trust game builds upon and extends Wilkinson-Ryan (2015)'s setup. Two players, *"principal"* ($P$) and *"agent"* ($A$), play sequentially and are endowed with six and zero tokens at the game's outset, respectively.[4] In the base game, both players make binary choices. First, player $P$ can opt into a transaction with $A$ by passing either four or zero tokens to $A$. If $P$ decides not to pass tokens to $A$, the game ends, and both players realize outside option payoffs, which correspond to their endowments. However, if $P$ passes four tokens to $A$, the four tokens triple such that $A$ obtains 12 tokens. $P$ has two tokens left. Next, player $A$ decides to pass back six or two tokens to $P$. If $A$ passes back six tokens, player $P$ earns eight tokens, and player $A$ earns six tokens. If $A$ passes back only two tokens, players $P$ and $A$ realize game payoffs of four and ten tokens, respectively.

As with many other versions of the trust game, this sequence of choices and the game's parametrization capture essential elements of contracts. First, $P$ and $A$ can realize a cooperative surplus. $P$ puts a valuable asset under $A$'s control before $A$ performs her part in completing the transaction. If $P$ does not initiate the transaction, the parties forego possible gains from trade. Therefore, rational parties would want to commit credibly to cooperating. However, $A$'s action is not contractible, i.e., the players cannot agree on incentive-compatible contracts ex-ante. Therefore, after $P$ committed, $A$ is incentivized to act opportunistically (moral hazard). $P$ may anticipate this behavior and, in turn, decide not to commit in the first place. Second, the exchange is deferred in the sense that completing the transaction takes time, i.e., players make

[4] I use the terms "principal" and "agent" here to adhere to the terminology of contract theory. I did not use those terms in the experiment because they carry hierarchical notions. Roles in the experiment were labeled "sender" and "receiver".

their decision sequentially. The temporal discrepancy between commitment and performance creates additional uncertainty regarding the choice environment.

Several features modify this basic game structure. First, the experiment features pre-play communication. Before $P$ decides whether to pass four tokens to $A$, the players can exchange messages in a computer chat. Communication spurs cooperation in social dilemmas (cf: Balliet, 2010) and features prominently in the literature about the effects of promises (e.g.: Charness and Dufwenberg, 2006; Vanberg, 2008; Ederer and Stremitzer, 2017) and in research on moral hazard problems (e.g.: Harbring, 2006; Hoppe and Schmitz, 2018). In this experiment, communication facilitates reducing strategic uncertainty and can thus encourage $P$ to initiate the transaction.

Moreover, once $P$ initiates the transaction but before $A$ decides how many tokens to pass back, the players learn that $A$ receives an opportunity to participate in another transaction (with the experimenter). Specifically, $A$ can pay 10 of her 12 tokens to work on a counting task. The counting task involves counting the numerical digit "0" in an $8 \times 25$ block of single numerical digits randomly drawn from "0" to "9". Player $A$ can earn between 10 and 41 tokens. The earnings in the counting task consist of a randomly determined base wage between 10 and 40 tokens and a bonus payment of one token if $A$ counts the correct number of zeros with an error margin of one. Participating in the counting task is a (weakly) dominant strategy for $A$ but requires an upfront payment of 10 tokens such that $A$ can pass back two tokens to $P$ at most. As in efficient breach settings in the real world, the additional counting task constitutes a spontaneously occurring temptation to devote resources towards the new transaction rather than the existing relationship.[5] Even to good-faith agents, who initially intended to cooperate, the new task provides cause to reconsider their choice.

As $A$ faces the choice between passing back six tokens or passing back only two tokens and paying ten tokens to participate in the counting task, I elicit her WTB with a random binary choice (RBC) mechanism. The RBC mechanism is procedurally identical to the Becker et al. (1964) mechanism and incentive-compatible under the lenient assumption that participants never choose dominated gambles (Azrieli et al., 2018). Participants receive a list with all possible payout combinations in the counting task, i.e., 10-11 tokens, 11-12 tokens, up until 40-41 tokens. For each possible payout combination, participants indicate their preference of either passing back six tokens to $P$ and keeping six tokens or passing back only two tokens to $P$ and paying ten tokens to participate in the counting task. The computer then randomly selects one payout combination. Next, $A$'s choice for that option is implemented. Each option is equally likely to be chosen. Instead of requiring participants to make each of the 40 binary choices separately, participants state their minimum required payout for the counting task, i.e., their reserve price, with a range slider. By positioning the slider, participants fill out the choice list accordingly.

While $A$ decides how many tokens she requires as a minimum earning in the counting task rather than passing back six tokens to $P$, player $P$ steps into the shoes of a hypothetical agent. $P$ receives the same information as $A$ about the nature of the counting task, the trade-off regarding the decision to pass back tokens, and the RBC elicitation procedure. The experiment instructs $P$ to think about the minimum guaranteed payout that she would require as a hypothetical agent

[5]The occurrence of the new transaction cannot have been all too surprising because participants so far only received information about the first stage of the multi-stage game. They were also told that they would receive information on later stages when and as soon as these occur. Section 2.3 contains detailed information on instructions and other procedures.

to prefer paying ten tokens to participate in the counting task. By filling out the same list of binary choices as $A$ does, $P$ thus submits her beliefs regarding $A$'s WTB. To encourage $P$ not to submit her WTB but rather her belief about $A$'s WTB, $P$ receives two tokens if she matches $A$'s minimum required payout with an error margin of three tokens. For instance, if the elicited WTB of $A$ were 29 tokens, $P$ would receive two tokens if she submitted a WTB-belief between 26 and 32 tokens.

Once $P$ submits her belief regarding $A$'s WTB, she proceeds to a donation stage. $P$ learns that she received a new budget of 100 cents, i.e., the equivalent of two tokens, and that she can decide how many cents to donate to $A$. The instructions for the donation round inform $P$ that the donation is "costless", i.e., $P$ will not retain the remaining amount. Employing the strategy method (Selten, 1967; Brandts and Charness, 2011), $P$ makes her donation decision for each possible choice of $A$, i.e., not accepting the counting task and passing back six tokens or participating in the counting task and passing back only two tokens. As $P$ does not keep the remaining amount, the difference between this budget and the donation constitutes a form of punishment. Therefore, the donation stage creates direct information about $P$'s disapproval of $A$'s actions (cf.: Depoorter and Tontrup, 2012).

Finally, after $P$ finishes the donation round and $A$ completes the WTB elicitation and, if applicable, the counting task, the players learn about the game outcomes and their payoffs. Afterward, they indicate to what extent they agree with three statements regarding their moral attitude towards participating in the counting task rather than sending back six tokens. The three statements are (see: Wilkinson-Ryan, 2015):

1. It is immoral to pay to participate in the counting task, because it leaves the Sender with less than he or she was expecting.

2. It is immoral to pay to participate in the counting task if the Receiver has agreed to pass back six tokens to the Sender, because it means going back on your word.

3. It is immoral to pay to participate in the counting task, because the Sender was generous and deserves reciprocal generosity.

Players indicate their agreement on a slider ranging from -50 to 50. These statements link participants' moral assessment to three prominent micro-foundations for pro-social behavior that may affect participants' WTB:

1. Guilt aversion has received considerable attention in the economic literature. The theory of guilt aversion holds that a person's perception of another person's expectations influences her decisions (e.g.: Charness and Dufwenberg, 2006; Bacharach et al., 2007; Battigalli and Dufwenberg, 2007; Dufwenberg et al., 2011; Beck et al., 2013; Hauge, 2016; Bellemare et al., 2017; Ederer and Stremitzer, 2017). Especially in the context of pre-play communication (Balafoutas and Sutter, 2017), when a person anticipates disappointing others relative to their expectations creates a psychological cost.

2. A preference for consistently following through on prior commitments can explain pro-social decisions. Once promisors make promises, they prefer to keep their word.

(e.g.: Ellingsen and Johannesson, 2004; Vanberg, 2008). The idea that commitment and consistency are essential drivers for cooperative behavior and that violating a prior commitment creates psychological costs also corresponds to a prominent research stream in social psychology (e.g.: Festinger, 1957; Cialdini, 2007).

3. Positive reciprocity can explain the cooperative decisions of agents. Positive and negative reciprocal behavior play a decisive role in many economic domains.[6] Seminal theories of reciprocal behavior model players' actions as a response to either the good intentions or the perceived kindness of others (cf.: Rabin, 1993; Levine, 1998; Charness and Rabin, 2002; Falk and Fischbacher, 2006). Given a preference for reciprocity, the intrinsic utility of the cooperative choice increases if *A* perceives the prior decision of *P* as well-intentioned or kind.[7]

Participants' moral assessment of opportunistic behavior was not incentivized. Moreover, the three behavioral channels discussed above are not mutually exclusive. The experiment's primary purpose is to serve as an institutional test bed. I do not aim to reliably disentangle the behavioral micro-foundations but rather the different components of compensatory damages.

## 2.2 Treatments

In the control condition ("baseline": BL), *A*'s opportunistic behavior does not trigger any additional monetary consequence for either player. The control condition implies that no legal rule for breach of contract exists. The control condition implies that no legal rule for breach of contract exists. The primary purpose of the experiment is to investigate the effect of liability for damages on *A*'s disposition to break a commitment. To this end, I conduct three more treatments that differ in the rule governing breach, i.e., there are different consequences for non-cooperative play.

As the central manipulation, treatment CO ("compensation") introduces a compensation mechanism as a breach remedy. If *A* passes only two tokens back to *P*, the computer automatically transfers four tokens from *A* to *P* at the end of the game.[8] The experiment tailors the amount of the compensation such that *P* is made whole, i.e., the compensation puts *P* in a situation as if *A* would have honored her commitment by sending six tokens. The compensation sanctions *A* for non-cooperative play and, simultaneously, insures *P* against potential opportunistic behavior of *A*. As long as *P* initiates the transaction, her payoff of eight tokens is independent of how *A* decides.

[6] See Malmendier et al. (2014) for an excellent recent review of economic research and evidence on reciprocal behavior.

[7] Malmendier et al. (2014) suggest and provide evidence that the seminal models fail to account for external motivations for cooperative behavior (e.g., social norms, social image, social signaling, audience effects, prestige, shame, guilt, and reputation). Consequently, prior studies on positive and negative reciprocity may substantially overestimate the effect of internal motives (e.g., fairness or altruism).

[8] In the real world, principals would have to claim damages instead of receiving the compensation payment automatically. Hoeppner et al. (2017) include a decision to claim compensation after the agent made her choices and nature revealed the outcome. The overwhelming majority of all participants who could claim compensation also did so. Therefore, I did not add another decision step in the present experiment.

As the compensation mechanism functions as both insurance and sanction, it affects the game payoff of both players. The transfer of tokens reduces *A*'s payoff and increases *P*'s payoff. To the extent that *A* considers both her own and *P*'s payoff when making her decision, i.e. when *A* exhibits social preferences, both components of the compensation mechanism can influence *A*'s WTB.

I conduct two additional treatments to disentangle the sanctioning effect from the insurance effect. In treatment FI ("fine"), *A* faces a fine instead of the obligation to compensate *P*. If *A* passes back only two tokens to *P*, a fine of four tokens will be deducted from her final game payoffs. The fine sanctions *A* for non-cooperative play but does not insure *P* against opportunistic behavior of *A*. Finally, treatment RE ("recovery") implements a recovery procedure. When *P* only receives two tokens from *A*, the experimenter will recover what *P* has lost due to *A*'s opportunistic choice, i.e., *P* will receive an extra four tokens. The recovery procedure does not sanction *A* for non-cooperative play. However, similar to the compensation mechanism, the recovery procedure insures *P* against potential opportunistic behavior of *A*. Therefore, in treatment RE *P*'s payoff from the transaction is again independent of how *A* decides.

In each treatment, the instructions inform all participants about the consequences of *A* choosing to send back two instead of six tokens.

## 2.3 Procedures & Sample

I programmed the experiment with oTree (Chen et al., 2016) and conducted it online, sampling participants from Amazon's Mechanical Turk (AMT). A demo version is available online.[9]

Participants navigating from the AMT task interface to the experiment first arrived at an informed consent form.[10] After providing informed consent, participants received general information about the experiment. This general information assured them that there was no deception. It also informed participants that the computer matches every participant with another participant and that the experiment does not employ robots. Participants also learned that their and others' decisions are directly related to their earnings and that two tokens in the experiment equal one US-Dollar. Finally, they learned that the experiment consists of three parts: (1) an initial socio-demographic questionnaire; (2) the main part featuring a multi-stage interactive game; and (3) a short exit survey.

The next step consisted of an attention check. Inattentive participants were excluded from the study. Participants who passed the attention check answered the initial socio-demographic questionnaire.

Afterward, participants received instructions about the trust game. They learned that the computer would pair them with another participant, they learned that the computer would randomly allocate them to play the trust game as either "Sender" or "Receiver", and they received

[9] The demo version exemplifies each treatment of the experiment as I conducted it. The demo version features only two players, which conveniently facilitates clicking through the experiment. Moreover, in the demo version, I incorporated minor changes in the code. These changes help migrate the experiment from AMT to the demo server. However, they do not change the appearance or feel of participating on screen. The demo version of the experiment is available online at: https://demo-cib.herokuapp.com/demo.

[10] The program checked the geolocation of participants and blocked all non-US participants and participants behind proxy servers. However, the program did not store the geolocation data. Moreover, the geolocation data was not accessible at any time. Participants were informed about this aspect on the informed consent form.

detailed information about each player's choices, the sequence of play, and the associated payoffs. Participants also learned that they would receive detailed information about their choices and payoffs in later stages as soon as these stages occurred. Finally, participants learned that they would have to answer control questions and be excluded from the study should they incorrectly answer the check questions.

Before starting the main part, participants received a short tutorial on the WTB elicitation procedure. I modeled the context of the tutorial after the one used by Wilkinson-Ryan (2015), adapting it to my different elicitation procedure. The tutorial aims to avoid possible confusion over the WTB elicitation method (cf.: Plott and Zeiler, 2005, 2011; Isoni et al., 2011). The tutorial describes a person who is interested in a job but does not know the pay. After explaining the person's reserve price, participants filled out a willingness-to-accept form on behalf of the person. Like the RBC mechanism described above, this form consists of a list of six possible payment options for the job, with a binary choice ("don't accept the job" or "accept the job") associated with each option. If participants did not fill out the form correctly, the program reminded them of the reserve price. The program required participants to fill out the form correctly. Next, the tutorial introduced participants to using a range slider to fill out the same form. Once participants correctly filled out the form, they learned what would happen if the job was worth more or less than the reserve price. Finally, the tutorial reminded participants that they do not have a strategic incentive to misreport their reserve price. Thus the experiment featured the recommended controls to ensure participants' comprehension of the WTB elicitation procedure (cf.: Plott and Zeiler, 2005; Predmore et al., 2021).

Next, participants arrived at a grouping stage where the computer matched with one other participant. Their roles were determined randomly. Participants then proceeded to the main part. Upon completion, they received detailed information about the game results and completed the exit survey. Once participants submitted the task on AMT, they could not participate in the experiment again.

A total number of 510 participants finished the experiment. Table 1 reports descriptive sample statistics by treatment.After conducting a power analysis, I intended to sample 48 observations per treatment, i.e., 96 players. Due to the AMT sampling procedure, the sample is unbalanced across treatments, i.e., different treatments have different numbers of participants. I monitored the number of submitted tasks on AMT and the number of dropped-out participants in the experiment and implemented the following stopping rule: for each treatment, cancel the task on AMT as soon as the number of submitted tasks on AMT minus the number of dropped-out participants equals 100.[11] However, when a requester cancels a task on AMT, it is merely removed from the list of available tasks that workers can browse. Participants who have already started the task can still finish, and workers who have opened the task's overview with the link to the experiment server can navigate to the experiment. Therefore, treatments exhibit a varying number of participants and principal-agent pairs.

[11] I set the threshold to 100 instead of $2 \times 48 = 96$ because workers can submit the HIT fraudulently, i.e., submit a made-up completion code. Live monitoring of all submitted completion codes versus correct completion codes appeared prohibitively tricky. Based on prior experience and suggestions by colleagues, I set a cushion of 4%.

**Table 1:** Overview of Sample Characteristics by Treatment

| | BL | CO | FI | RE | Total |
|---|---|---|---|---|---|
| N | 132 | 110 | 108 | 160 | 510 |
| female (share) | 49 (0.371) | 45 (0.409) | 41 (0.380) | 57 (0.356) | 192 (0.377) |
| age: | | | | | |
| median | 34 | 39 | 35 | 35 | 35 |
| st. dev. | 10.472 | 11.438 | 9.999 | 9.924 | 10.499 |
| AMT hours per week: | | | | | |
| median | 20 | 20 | 20 | 20 | 20 |
| st. dev. | 14.698 | 13.182 | 12.892 | 10.938 | 13.089 |
| median or above household income (share) | 30 (0.227) | 23 (0.218) | 45 (0.213) | 57 (0.281) | 122 (0.239) |
| college degree and higher (share) | 88 (0.667) | 76 (0.691) | 71 (0.658) | 117 (0.731) | 352 (0.690) |
| employed or freelance (share) | 119 (0.902) | 97 (0.882) | 91 (0.843) | 141 (0.881) | 352 (0.878) |
| ethnicity: caucasian (share) | 99 (0.750) | 89 (0.809) | 84 (0.778) | 105 (0.656) | 448 (0.739) |

At the end of the experiment, tokens earned by participants were exchanged to US Dollars with an exchange rate of 2:1. On average, participants earned $9.56, including a completion fee of $3.00.[12] The average participant submitted the task on AMT after 42 minutes and 3 seconds.

## 3 Behavioral hypotheses

To derive predictions about agents' decisions in the experiment, I dissect the compensation mechanism into two parts: (1) the amount the agent pays and (2) the amount the principal receives, i.e., fine and recovery. The idea is that fine and recovery can have distinct behavioral effects, which coincide under a compensation regime.

The fine takes center stage in the standard economic theory of remedies. Remedies in contract law, such as compensatory damages, ought to turn games with non-cooperative solutions into games with cooperative solutions (Cooter and Ulen, 2016, Ch. 4). Absent more valuable outside options at the time of entering the contract, sufficient monetary costs created through remedial mechanisms make contract performance the preferred choice for the agent, render her initial commitment credible, and thus align incentives of the parties.[13] Classical economic theory would predict that the increased monetary costs for non-performance reduce the agent's WTB. Classical theory, however, does not account for non-monetary elements in the agent's utility function, which can interact with monetary elements. For instance, moral, psychological, or otherwise subjective costs may trade off the monetary costs created through contract remedies.

A prominent explanation of why the extra monetary burden of contract remedies may tip the scales between monetary and psychological costs evokes social norms. In economics, social norms have also received increasing attention as determinants for pro-social behavior (Bicchieri, 2006; Malmendier et al., 2014). Specifically, Kessler and Leider (2012) propose that agreements establish norms endogenous to the relationship that agents feel obliged to follow, despite being cheap talk. In technical terms, to the extent that their actions deviate from the established social

[12]Participants whose assigned partner dropped out received the average earnings as payment.

[13]Parties to a contract can set remedies for breach, i.e., party-stipulated damages themselves. If these are stipulated unilaterally, however, there is a risk of negative reciprocity that increases ex-post opportunistic behavior of agents (Hoeppner et al., 2017).

norm, norm-sensitive parties to an agreement experience disutility. However, contract remedies such as compensatory damages may provide contextual cues that re-frame the underlying norms of the relationship. Starting with the much-discussed study by Gneezy and Rustichini (2000), a broad stream in the literature documents that, following the imposition of payment, individuals may re-interpret what is appropriate behavior (e.g. Brekke et al., 2013; Fehr and Rockenbach, 2003; Fehr and List, 2004; Mellström and Johannesson, 2008; Ariely et al., 2009).[14] What otherwise constitutes inappropriate behavior or an obligation to omit some action—e.g., breach of contract—may be perceived as permitted (for a payment) once altering the institutional environment provides new meaning to the underlying social norm. Consequently, agents may experience less disutility from breach of contract because costly nonperformance does not conflict with the underlying social norm. I hypothesize:

> **Hypothesis 1:** The agent's reserve price for participating in the counting task will be lower in treatment FI than in treatment BL.

In contrast to the fine element, the recovery element of the compensation mechanism has received considerably less attention. This lack of interest is not astonishing. After all, in classical economic theory, the principal's payoff does not enter the agent's utility function. In behavioral economics, however, research on social preferences, i.e., preference structures that also include the well-being of others in one way or another, figures prominently. Three phenomena from that area may affect the agent's breach decision.

First, contract remedies may affect behavior, notably through different specifications of preferences for reciprocity. (e.g.: Falk and Fischbacher, 2006; Cox et al., 2007). These intention-based models assume that persons form beliefs about others' underlying intentions of an action. The perceived kindness of an action depends on what alternative actions are available to the other person and on a person's beliefs about what the other will do, to the extent that these beliefs carry information about the agent's intention. In treatment RE, principals are insured by the recovery mechanism. Therefore, agents may have difficulty interpreting the principal's act of initiating the transaction as genuinely kind because there are no monetary consequences for the principal. As a result, agents have no unambiguous signal to evaluate the principal's intention. Consequently, there is no obvious trigger point for positive intention-based reciprocity in treatment RE compared to treatment BL.

Second, social preferences in the form of inequality or inequity aversion (see: Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) may alter the agent's breach decision. These outcome-based models assert that persons care about others' payoffs relative to their payoffs. The recovery mechanism reduces the inequality of outcomes.[15] To the extent that agents are inequality averse, agents experience less disutility from nonperformance in treatment RE than in treatment BL.

[14] Kornhauser et al. (2020) provide a splendid review of the literature. Two interesting replication attempts, a laboratory experiment (Kornhauser et al., 2020) and a vignette study (Metcalf et al., 2020), find mixed evidence for and qualify the a-fine-is-a-price hypothesis of Gneezy and Rustichini (2000).

[15] So do fines. However, the fact that the agent also loses a direct monetary benefit counterbalances the positive effect of less unequal outcomes. Therefore, I did not mention inequality and inequity aversion when discussing fines above.

Third, a stream of experimental evidence suggests a behavioral force of promise-keeping. Although statements of intent are mere cheap talk, exchanged promises and, thus, agreements enhance cooperation (e.g.: Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006; Vanberg, 2008; Sutter, 2009). Ederer and Stremitzer (2017) disentangle two leading explanations for the behavioral effect of promissory commitment. One explanation is that promises move beliefs about the promisee's expectations. Guilt-averse promisors suffer disutility from disappointing the promisee's expectations, even more so if these expectations were induced or strengthened by the promise. As a consequence, promisors are more likely to keep their promises in order to avoid guilt (Charness and Dufwenberg, 2006; Ederer and Stremitzer, 2017). The other explanation is that promisors have a belief-independent preference for promise-keeping per se (e.g.: Ellingsen and Johannesson, 2004; Vanberg, 2008; Ismayilov and Potters, 2012). This preference could, for instance, be modeled as a psychological cost of being inconsistent (Ellingsen and Johannesson, 2004) or a psychological cost of lying (Chen et al., 2008; Kartik, 2009).[16] These explanations for the behavioral force of promises and agreements are not exclusive (Mischkowski et al., 2019). However, to the extent that the recovery mechanism stabilizes the promisee's expectations about the transaction's outcome, it also mutes the promisor's guilt aversion. The recovery aspect of compensatory damages thus countervails some share of the behavioral effect of promissory commitment.

As the three aforementioned behavioral channels regarding the recovery mechanism work in the same direction, I hypothesize:

> **Hypothesis 2:** The agent's reserve price for participating in the counting task will be lower in treatment RE than in treatment BL. (Note that this prediction partly conditions on a perceived agreement in case of guilt aversion.)

In combination, fine and recovery mechanisms provide compensatory damages. The predicted effects of both aspects point in the same direction. Therefore, under compensation damages, I also expect a higher WTB, i.e., a lower reserve price for breach. Whether the isolated effects substitute or complement each other is an empirical question. I merely hypothesize:

> **Hypothesis 3:** The agent's reserve price for participating in the counting task will be lower in treatment CO compared to treatment BL.

In addition to the effect of remedial mechanisms on the agent's propensity to breach, remedies may also affect the extent to which the principal evaluates the breach decision as more or less reproachable. This idea speaks to the indemnification and vindication function of compensatory damages. Concerning the punishment measure, I predict that the non-cooperative decisions of agents are less reproachable to principals in treatments CO and RE because the principal does not suffer any monetary consequences. I hypothesize:

[16] This explanation also links promissory commitment to topics of consistent behavior in social psychology (cf.: Festinger, 1957; Cialdini, 2007).

**Hypothesis 4:** Principals will punish the agents less for non-cooperative choices in treatment CO and RE than treatment BL.

However, once the principal and agent have agreed upon exchange, the principal's perceptions of the agent's non-cooperative action may flip. After explicit agreement to cooperation, non-cooperative behavior is much more culpable—especially when the consequences of doing so emphasize the culpability through monetary loss in treatments CO and FI. Consequently, I predict:

**Hypothesis 5:** Perceiving the transaction as an agreement increases the punishment of the non-cooperative choices of the agent. The increased punishment given perceived agreement is more substantial in treatments CO and FI than in BL.

While the choice data do not facilitate isolating the different behavioral channels triggering the effects of remedial mechanisms on breach and punishment decisions, participants' answers in the exit survey may provide credence for some explanation over another. Therefore, my last set of predictions concerns participants' agreement with the statements about participating in the counting task as immoral.

The statements in the exit survey offered three reasons why participating in the counting task instead of sending back six tokens can be morally blameful. The first statement assessed the immoral nature of participating in the counting task due to disappointing the principal's expectations. In treatments CO and RE, the principal's expectations about their payoffs from the transaction are independent of the agent's decision. Therefore, as long as the principal's expectations do not condition on how her payoff is created, the agent cannot disappoint the principal's expectations. Therefore, I predict:

**Hypothesis 6:** Participants rate statement 1 (disappointing the principal's expectations) as less important in treatments CO and RE than in treatments BL and FI.

The second statement claims that participating in the counting task is immoral because the agent who had initially agreed to pass back six tokens would go back on her word. Evaluating such reneging as morally questionable implies a preference for consistently following through with a prior commitment. Note that this statement conditions on the agent's perception of an actual agreement. Whether and how the treatments affect the value of consistently following through with a prior commitment conditional on a perceived agreement is not immediately apparent. Therefore, I remain skeptical and predict:

**Hypothesis 7:** Given prior agreement, neither of the treatments CO, FI, and RE will increase or decrease the importance rating of statement 2 (going back on one's word) relative to treatment BL.

Finally, the third statement proposes that participating in the counting task is immoral because the principal was generous and deserves reciprocal generosity. In treatments CO and

RE, I expect less reciprocal kindness because the principal is insured. As there is no apparent trigger point for positive intention-based reciprocity in treatment CO and RE, I hypothesize:

> **Hypothesis 8:** Participants rate statement 3 (the principal deserving reciprocal generosity) as less important in treatments CO and RE than in treatments BL and FI.

# 4 Results

In this Section, after a methodological note, I first focus on treatment differences on the aggregate level, i.e., the society of principals and agents. Next, I zoom in and analyze the results on the individual level.

## 4.1 Prelimenary note on inference

### 4.1.1 A Bayesian approach is preferable for this study

I employ a Bayesian approach throughout. Bayesian statistics is becoming more and more common in experimental social and behavioral research. The profound reproducibility crisis in psychology (Open Science Collaboration, 2012, 2015) and similar problems in other fields, including economics (Camerer et al., 2016; Ioannidis et al., 2017), possibly drive the increased advocacy for Bayesian approaches. Concomitant calls for rethinking statistical inference from data exist (e.g.: Dienes, 2011; Wagenmakers et al., 2011; Colling and Szűcs, 2018).

Using a Bayesian approach for statistical inference, I do not dogmatically argue its superiority over frequentist methods. Bayesian approaches have distinct advantages and some disadvantages.

In the present context, two key characteristics of the data generating process advise employing Bayesian rather than frequentist methods.[17] First, I could not strictly enforce the stopping rule for data collection, i.e., triggering the stopping rule did not automatically prevent further data collection. When I canceled the task on AMT according to the stopping rule, participants who had already started could still finish the experiment. Moreover, AMT workers who had already navigated to the task overview could still start the experiment. I prefer keeping the additionally sampled data because it carries relevant information and increases power. Second, I never planned treatment RE a priori. Instead, collecting data for treatments BL, CO, and FI made the necessity of treatment RE apparent.

For these two reasons, experimenter intentions changed during data collection: namely, intentions regarding (1) sampling procedures, (2) the number of conditions, and (3) the comparison of collected data with any other condition. In frequentist approaches, however, the step from the test statistic to the p-value crucially depends on experimenter intentions (Kruschke, 2010). By contrast, Bayesian inference does not rely on these and similar intentions. It only conditions on the observed data and, therefore, facilitates overcoming these concerns for inference.

[17] Nor do I find it surprising or discouraging that a substantial share of scientific results across fields proves elusive to replication. After all, small samples are noisy, and human participants are heterogeneous. Therefore, replication failure is typical, and experimenting multiple times should be accepted as a standard—and a publishable—part of scientific work.

The online supplementary material contains the frequentist analog to the Bayesian analysis. All main results prevail under the frequentist approach.[18]

### 4.1.2 Decision rules to evaluate posterior parameter distributions

Bayesian inference provides a posterior distribution of credible parameter values from a prior distribution and the available data. Parameter values more consistent with the data receive a higher probability than those less consistent. I will summarize a given posterior parameter distribution with its mode, i.e., the maximum a posteriori probability estimate (MAP), and its 95% highest density interval, i.e., the 95% of the most credible parameter values (0.95-HDI).

To evaluate the results, I conceptually distinguish between the existence of an effect and its credibility (see: Makowski et al., 2019). To assess the existence of an effect, I analyze how much of the posterior's probability mass $P(\theta)$ covers the positive or negative domain and whether 0.95-HDI includes zero. Therefore, the threshold to acknowledge the existence of an effect is 0.975 of the posterior probability mass. I gauge an effect's credibility by comparing the entire posterior distribution with a region of practical equivalence, i.e., ROPE (cf.: Kruschke, 2015, 2018; Lakens et al., 2018; Makowski et al., 2019). ROPE specifies a range of parameter values that I consider equivalent to the null value. For linear models, I set ROPE bounds to $\pm 0.1 \times \sigma_y$, where $\sigma_y$ is the standard deviation of the dependent variable. I set ROPE bounds to $\pm 0.181$ when models express parameter values in log odds ratios. When I re-scale independent variables, I adjust ROPE bounds using the scaling parameters. I define the probability mass of the posterior that overlaps with ROPE as $P_{ROPE}$. If at least 97.5% of the posterior lies within ROPE, i.e., $P_{ROPE} \geqslant 0.975$, I accept the null value. If at most 2.5% of the posterior probability mass lies within ROPE, i.e., $P_{ROPE} \leqslant 0.025$, I reject the null value and consider the effect credible. If neither, I withhold judgment.[19]

## 4.2 Aggregate willingness-to-breach

The lower panel of Table 2 summarizes the relative frequency of non-cooperative choices, i.e., agents who kept ten tokens and sent back two tokens to principals. Overall, agents participated in the counting task 54% of the time instead of cooperating with principals. Absent a remedy, the relative frequency of opportunistic choices is 46.30% in treatment BL. For principals who receive compensation from agents in treatment CO or recovery in treatment RE, this relative frequency increases to 67.27% and 60.53%, respectively. When agents have to pay a fine for engaging in the counting task in treatment FI, however, the relative frequency of sending back two tokens decreases to 38.46%. It appears as if compensation and recovery have a positive effect on non-cooperative behavior, whereas fines reduce non-cooperative choices. Note, however,

[18] The supplementary material can be retrieved from the author or the paper's website at: Insert corresponding URL here. [For the purpose of referee review, an extra appendix reports the frequentist analysis.]

[19] Note that the choice of criteria for effect existence and credibility are arbitrary but resemble sensible default values used in the literature. Nevertheless, the thresholds employed here are no make-or-break decision criteria. Rather they represent one particular value of continuous indices of effect existence and credibility. Moreover, I avoid an entirely different index for the evidence, namely Bayes Factors. I refrain from reporting Bayes Factors not because Bayes Factors are uninteresting but because I do not want to overburden the report with yet another statistic. However, I strongly encourage readers to explore the data further.

that the implementation of the agents' individual decisions condition on a random draw in the WTB elicitation procedure.

**Table 2:** Relative Frequencies of sending choices and opportunistic choices, conditional on treatment and perceived agreement.

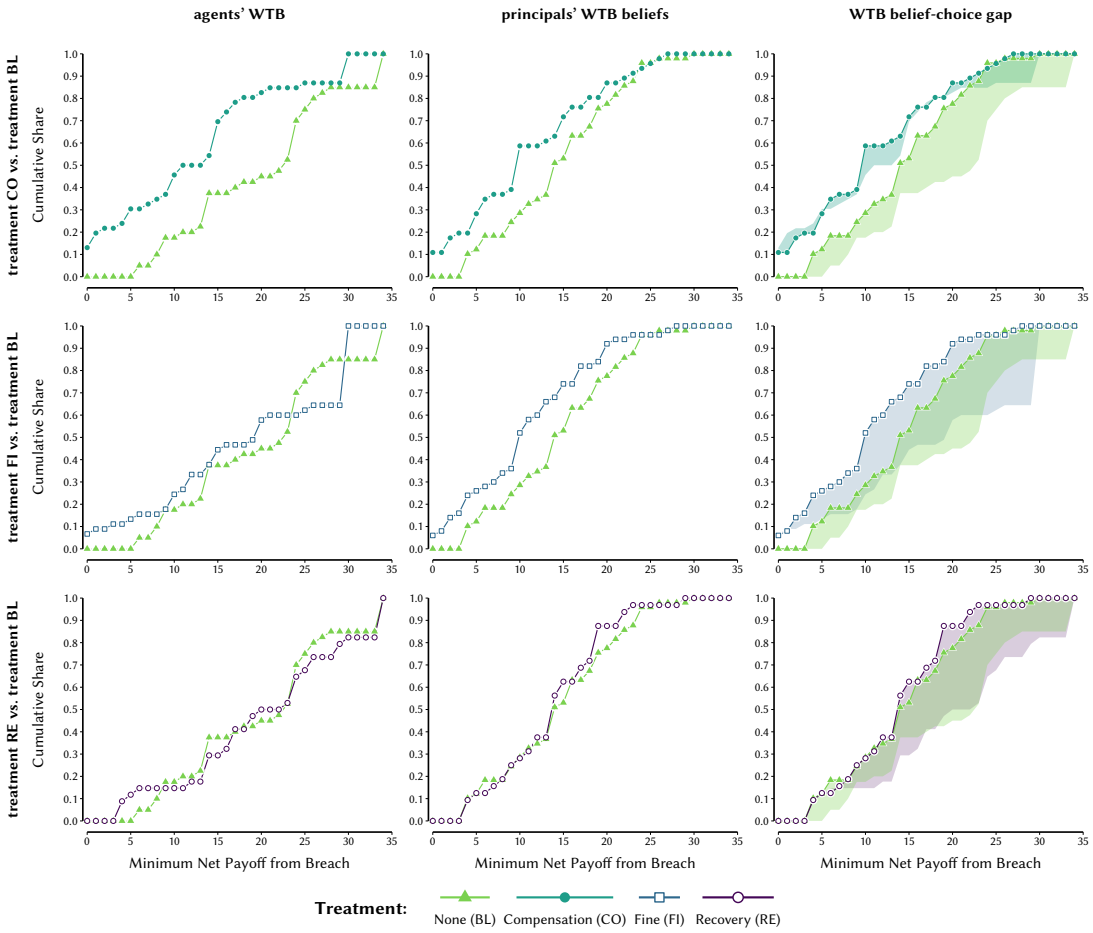
| | | Treatment | | | | |
|---|---|---|---|---|---|---|
| principal initiates (sends 4 tokens) | N | BL | CO | FI | RE | Total |
| all | 255 | 0.818 | 1.000 | 0.963 | 0.950 | 0.929 |
| perceived agreement | 180 | 0.980 | 1.000 | 0.980 | 0.970 | 0.983 |
| perceived disagreement | 75 | 0.313 | 1.000 | 0.667 | 0.936 | 0.800 |
| agent breaches (sends back 2 tokens) | | | | | | |
| all | 237 | 0.463 | 0.673 | 0.385 | 0.605 | 0.540 |
| perceived agreement | 165 | 0.525 | 0.630 | 0.333 | 0.382 | 0.473 |
| perceived disagreement | 72 | 0.286 | 0.889 | 0.714 | 0.786 | 0.694 |

Rather than actual opportunistic behavior, my main hypotheses concern how compensation affects agents' WTB. The left column of Figure 1 depicts the cumulative share of agents opting for the counting task as a function of net profit, i.e., the aggregate WTB. Each panel compares one of the treatments CO, FI, or RE, with treatment BL. In the control treatment BL, agents preferred passing back six tokens to the principal over the counting task unless the net profit from opportunistic behavior was at least six tokens. Moreover, half of the agents would not act opportunistically unless net profits reached 23 tokens. Finally, all agents would choose to work on the counting task in treatment BL only at the highest possible net profit of 34 tokens. The picture is very different for treatment CO. Even when choosing the counting task does not yield net profit, 13% of agents already decide opportunistically. At a net profit of 11 tokens, 50% of agents prefer the counting task over sending back six tokens to principals. That is less than half of the token required in the control treatment. All agents in treatment CO prefer the opportunistic over the cooperative choice for a net profit of 30 tokens. Compared to treatment BL, the aggregate WTB also increases in treatment FI. At a net profit of 0 tokens, 6.67% of agents prefer to work on the counting task. Below 20 tokens, the share of agents willing to forego cooperative play does not exceed 50%. Like treatment CO, the aggregate WTB in treatment FI caps out at 30 tokens. Differences in aggregate WTB between treatments RE and BL, by contrast, are difficult to discern visually. The initial threshold to induce non-cooperative play is somewhat lower than in treatment BL, namely four tokens instead of six tokens. Like treatment BL, half of the agents would not act opportunistically unless net profits reached 22 tokens, and only at the maximum net profit of 34 tokens do all agents in treatment RE prefer the counting task over passing back six tokens to principals.

I quantify these aggregate results by employing Ferrari and Cribari-Neto (2004)'s beta regression framework. Model M1 estimates the effects of treatment dummies for CO, FI, and RE. Model M2 also estimates net profit's effect and the interactions with the treatment variables. Following Gelman (2008), I scale the additional predictors to facilitate a comparison of the effects to the effects of the treatment dummies.[20] The left column of Figure 2 reports the results

[20]All prior distributions used throughout the analysis are minimally informative. The simulations converge very

**Figure 1:** Cumulative shares of agents' willingness-to-breach (left column), of principals' beliefs about agents' willingness-to-breach (middle column), and of the difference between the former and the latter (right column), all as a function of net profits. Each row represents one treatment comparison between treatments CO, FI, and RE with the control treatment BL.




by depicting the posterior distributions of the model parameters. When estimating treatment effects only (M1), I find evidence for positive effects of treatments CO and FI on aggregate WTB. Given the observed data, the effect of treatment CO has a probability of 1.00 of being positive (MAP $= 1.441$, 0.95-HDI [0.831, 2.057]) and is credible ($P_{ROPE} < 0.001$). Likewise, the effect of treatment FI has a probability of 1.00 of being positive (MAP $= 1.158$, 0.95-HDI [0.573, 1.813]) and is credible ($P_{ROPE} = 0.007$). When principals recover their loss from opportunistic behavior in treatment RE, by contrast, the probability mass of the posterior parameter distribution in the

well to the posterior distributions. The largest potential scale reduction factor among the simulated parameters is close to one, i.e., $\max(\hat{R}) < 1.001$ for M1 and $\max(\hat{R}) < 1.001$ for M2. Among all parameters, the smallest ratio of the effective sample size to the number of draws is $\min(n_{eff}/N) = 0.915$ for M1 and $\min(n_{eff}/N) = 0.935$ for M2. To avoid clutter, I do not depict MCMC diagnostic plots. Interested readers can consult the corresponding R-script in the supplementary material available from the article's website or the author.

positive domain is only 0.698 (MAP = 0.197, 0.95-HDI [-0.462, 0.729]).

These results remain robust when adding net profit and its interaction with the treatment indicators to the estimation (M2). The posterior parameter distributions of the treatment dummies in M1 and M2 strongly overlap. If anything, in M2 the effect of treatment CO gets a little bit stronger (MAP = 1.861, 0.95-HDI [1.402, 2.196], $P_{ROPE} = 0$), and the effect of treatment FI gets a little bit weaker (MAP = 1.028, 0.95-HDI [0.612, 1.411], $P_{ROPE} = 0$). Moreover, the effect of net profit is credibly positive (MAP = 4.532, 0.95-HDI [3.978, 5.120], $P_{ROPE} = 0$). The interaction effect between net profit and treatment CO has a probability of 0.988 of being negative (MAP = −0.891, 0.95-HDI [-1.621, -0.118]). However, as I adjust ROPE due to the re-scaling of the predictor, the negative effect is not credible. In fact, the results confirm the null hypothesis for practical considerations ($P_{ROPE} = 0.993$). The estimation also yields evidence for null effects for the interaction of net profit with treatment FI (MAP = −0.369, 0.95-HDI [-1.133, 0.421], $P_{ROPE} > 0.999$) and with treatment RE (MAP = −0.412, 0.95-HDI [-1.172, 0.356], $P_{ROPE} > 0.999$).

> **Result:** On the aggregate level, agents display an increased willingness-to-breach when a compensation remedy or a fine mechanism exists. A recovery procedure, by contrast, has no discernible effect on aggregate willingness-to-breach.

Next, I investigate aggregate principals' beliefs about their agent's WTB, which yields similar findings. The middle column of Figure 1 indicates elevated aggregate WTB beliefs in treatments CO and FI compared to treatment BL. By contrast, aggregate WTB beliefs in treatment RE visually appear very close to aggregate WTB beliefs in control treatment BL. I again use a beta model and the same sets of predictors to estimate the effects.[21] The middle column of Figure 2 visually reports the results. When estimating the effects of the treatment indicators only, I do not find convincing evidence for increased aggregate beliefs of principals about agents' WTB. Treatments CO (MAP = 0.558, 0.95-HDI [-0.009, 1.220]) and FI (MAP = 0.509, 0.95-HDI [-0.076, 1.145]), respectively, have a probability of 0.970 and 0.955, respectively, to be positive. Although the effects are borderline non-existent, too much probability mass resides within ROPE in treatments CO ($P_{ROPE} = 0.093$) and FI ($P_{ROPE} = 0.117$). Treatment RE does not affect aggregate beliefs about WTB (MAP = 0.095, 0.95-HDI [-0.519 0.725].

The results change in the second estimation, which accounts for net profit and its interactions with the treatment variables. The effect of treatment CO has a probability of 1 to be positive (MAP = 1.033, 0.95-HDI [0.717, 1.350]) and is credible ($P_{ROPE} = 0$). Likewise, the entire posterior parameter distribution of treatment FI lies in the positive domain (MAP = 0.990, 0.95-HDI [0.673, 1.318]) beyond ROPE boundaries ($P_{ROPE} = 0$). By contrast and similar to agents' WTB choices, the effect of treatment RE on principals' aggregate WTB beliefs is neither existing (MAP = 0.237, 0.95-HDI [-0.097, 0.567]) nor credible ($P_{ROPE} = 0.358$). As the results in the middle column of Figure 2 indicate, the effect of net profit has a probability of 1 to be positive (MAP = 6.162, 0.95-HDI [5.672, 6.769]) and is credible ($P_{ROPE} = 0$). Although negative interaction effects exist

[21] The simulations converge very well to the posterior distributions. The largest potential scale reduction factor among the simulated parameters is close to one, i.e., $\max(\hat{R}) < 1.001$ for M1 and $\max(\hat{R}) < 1.001$ for M2. Among all parameters, the smallest ratio of the effective sample size to the number of draws is $\min(n_{eff}/N) = 0.929$ for M1 and $\min(n_{eff}/N) = 0.833$ for M2.

between net profit and treatment CO (MAP $= -1.456$, 0.95-HDI [-2.194, -0.790]) and net profit and treatment FI (MAP $= -1.338$, 0.95-HDI [-2.108, -0.690]), they are not credible ($P_{ROPE} = 0.853$ for CO and $P_{ROPE} = 0.899$ for FI). The interaction between net profit and treatment RE only has a probability of 0.862 of being positive (MAP $= 0.410$, 0.95-HDI [-0.319, 1.168]), and the estimation confirms a null result for practical purposes ($P_{ROPE} > 0.999$). To sum up, after accounting for net profits and interactions of net profits with the treatment variables, the analysis yields convincing evidence for increased aggregated beliefs of principals about agents' WTB in treatments CO and FI compared to BL.

> **Result:** Compensation and fine mechanisms elevate principals' aggregate beliefs about agents' willingness-to-breach. The recovery procedure, by contrast, does not affect principals' aggregate beliefs about agents' willingness-to-breach.

Finally, I examine the relation between principals' aggregate beliefs about agents' WTB and aggregate agents' WTB by computing the gap between the two. The right column of Figure 1 visualizes the treatment comparisons and illustrates two main insights about this WTB belief-choice gap. First, the WTB belief-choice gap in treatment CO is narrow compared to treatments BL, FI, and RE. Second, while the WTB belief-choice gaps in treatments FI and RE overlap strongly with the WTB belief-choice gap in treatment BL, the WTB belief-choice gaps of treatment CO and treatment BL coincide very little.

Estimating a linear model lends credence to this observation. I standardize the difference between principals' aggregate WTB beliefs about agents' WTB but otherwise estimate the same model structures as before.[22] The right column of Figure 2 reports the estimation results, which support the visual finding. When estimating treatment effects only (M1), the effect of treatment CO on the standardized WTB belief-choice gap has a probability of 1.00 to be negative (MAP $= -0.935$, 0.95-HDI [-1.324, -0.521]) and is credible ($P_{ROPE} < 0.001$). The effect of treatment FI has a probability of 0.976 to be positive (MAP $= 0.359$, 0.95-HDI [0.024, 0.835]), but too much of the posterior probability mass falls within ROPE to qualify as credible ($P_{ROPE} = 0.064$). The posterior parameter distribution of treatment RE indicates that an effect does not exist (MAP $= 0.233$, 0.95-HDI [-0.204, 0.610]). These results are robust to including the additional predictors (M2). The treatment effects do not change appreciably. While the effect of net profits has a probability of 0.988 to be positive (MAP $= 0.613$, 0.95-HDI [0.098, 1.190]), the effect is not credible given the scaled ROPE bounds ($P_{ROPE} = 0.913$). The interaction effects of net profit with the treatment variables are neither existent nor credible. Moreover, the differences between treatments CO and FI (MAP $= -1.374$, 0.95-HDI [-1.704, -0.946], $P_{ROPE} = 0$) and between treatments CO and RE (MAP $= -1.108$, 0.95-HDI [-1.511, -0.756], $P_{ROPE} = 0$) are also credibly negative. In sum, treatment CO reduces the difference between aggregate beliefs of principals about agents' WTB and agents' WTB.

[22] The simulations converge very well to the posterior distributions. The largest potential scale reduction factor among the simulated parameters is close to one, i.e., $\max(\hat{R}) < 1.001$ for M1 and $\max(\hat{R}) < 1.001$ for M2. Among all parameters, the smallest ratio of the effective sample size to the number of draws is $\min(n_{eff}/N) = 0.922$ for M1 and $\min(n_{eff}/N) = 0.945$ for M2.

**Result:** Under a compensation mechanism, aggregate principals' beliefs about agents' WTB are much closer to aggregate agents' WTB than absent compensation or under fine or recovery mechanisms.

## 4.3 Individual-level analysis

### 4.3.1 Principals' choices to initiate transactions

I start the individual-level analysis by investigating the choices of principals to initiate the transaction. The upper panel of Table 2 reports the share of principals who sent four tokens to their corresponding agent. In the baseline treatment, 81.81% of all principals send four tokens to the agent. This share increases to 100% in treatment CO, 96.30% in treatment FI, and 95.00% in treatment RE. I estimate a Bayesian linear probability model to quantify the effects of treatments CO, FI, and RE on principals' choices to initiate the transaction.[23] The left column in Figure 3 depicts the posterior distributions of the model parameters, which explicitly illustrates the relative credibility of the parameter values. The results present convincing evidence that principals in treatments CO, FI, and RE are more likely to send four tokens to the agent than principals in the control treatment BL (M1 in Figure 3). The effect of treatment CO is credibly positive. The posterior parameter estimates have a probability of 1.00 to be positive (MAP $= 0.172$, 0.95-HDI [0.089,0.267]) and are credible ($P_{\text{ROPE}} < 0.001$). Similar results hold for treatments FI and RE. The posterior parameter distribution of FI has a probability of 0.999 of being positive (MAP $= 0.144$, 0.95-HDI [0.055, 0.233]), and the 0.95-HDI does not overlap with ROPE ($P_{\text{ROPE}} = 0.006$). The posterior parameter distribution of RE has a probability of 0.999 of being positive (MAP $= 0.135$, 0.95-HDI [0.050, 0.214]) and is credible ($P_{\text{ROPE}} = 0.005$).

**Result:** Compensation, fine, and recovery remedies increase the trusting behavior of principals.

This result becomes more differentiated when considering whether a principal evaluates the communication with the agent as an agreement to exchange tokens cooperatively. When principals perceive such an agreement, the relative frequency of initiated transactions remains similar across treatments. Specifically, the share of principals who sent four tokens to the agent is 98.00% in treatment BL, 100.00% in treatment CO, 98.04% in treatment FI, and 96.97% in treatment RE (Table 2). However, when principals do not evaluate the communication with agents as an agreement for cooperative exchange, a steep increase in the relative frequency occurs in treatments CO, RE, and FI compared to treatment BL. Specifically, the relative frequency of initiated transactions increases from 31.25% in treatment BL to 100.00%, 66.67%, and 93.62% in treatments CO, RE, and FI, respectively.[24]

[23] The simulation converges very well to the posterior distribution. The largest potential scale reduction factor among the simulated parameters is close to one, $\max(\hat{R}) < 1.001$. Among all parameters, the smallest ratio of the effective sample size to the number of draws is $\min(n_{\text{eff}}/N) = 0.934$.

[24] A cautionary note is in order. For perceived disagreement, the number of observations is relatively small in treatments CO and FI ($N_{\text{CO|disagree}} = 9$, $N_{\text{FI|disagree}} = 3$). Relying on asymptotic assumptions may not be diligent

I estimate the same linear probability model as before but now add predictors for perceived agreement and the interactions between perceived agreement and treatments CO, FI, and RE (M2 in Figure 3).[25] Absent a perceived agreement, the treatment effects of CO, FI, and RE relative to BL are credibly positive. The effect of treatment FI is the weakest of the three. Nevertheless, the posterior parameter distribution has a probability of 0.997 of being positive (MAP $= 0.365$, 0.95-HDI [0.114, 0.613]), and the 0.95-HDI does not overlap with ROPE ($P_{ROPE} = 0.004$). Perceiving an agreement also has a strong and credible positive effect on the decision to send four tokens. The entire probability mass of the posterior lies in the positive domain beyond ROPE (MAP $= 0.665$, 0.95-HDI [0.558, 0.788], $P_{ROPE} = 0.000$). Interestingly, the interactions between perceived agreement and treatments CO, FI, and RE are credibly negative. The interaction of perceived agreement and treatment FI is the weakest of the three interaction effects. Still, its posterior parameter distribution has a probability of 0.995 of being negative (MAP $= -0.351$, 0.95-HDI [-0.623, -0.101]) and is credible ($P_{ROPE} = 0.005$). To sum up, treatments CO, FI, and RE have a credible positive effect on the tendency of principals to initiate the transaction relative to BL absent perceived agreement. Conversely, treatments CO, FI, and RE have a credible negative credible effect on the likelihood given perceived agreement. I had no predictions regarding this auxiliary finding.

> **Auxiliary findings:** (1) Perceived agreement positively affects the initiation of transactions. (2) Absent perceived agreement, all external remedies (compensation, fine, and recovery) positively affect the initiation of transactions. (3) External remedies and perceived agreement have a negative interaction effect on the initiation of transactions.

The results do not appreciably change when adding controls for age, gender, education, and experience in online labor markets to the model (M3 in Figure 3).[26] The posterior parameter distributions of M2 and M3 overlap considerably, which is also illustrated by the similar 0.95-HDIs. The effects are robust to the controls.

### 4.3.2 Agents' individual WTB

I now turn to analyze agents' individual WTB, i.e., agents' reserve price for participating in the counting task. I standardize the reserve price measure and estimate the same model structure as before. One linear model (M1) quantifies the effects of treatments CO, FI, and RE on the individual WTB. A second model (M2) additionally considers agents' perception of having an agreement with the principal and the interaction effects of that perception with the

in this situation. However, removing the need for these assumptions with a Bayesian approach comes at the cost of possible sensitivity to priors.

[25] The simulation converges very well to the posterior distribution. The largest potential scale reduction factor among the simulated parameters is close to one, $\max(\hat{R}) < 1.001$. Among all parameters, the smallest ratio of the effective sample size to the number of draws is $\min(n_{eff}/N) = 0.978$.

[26] The control variable for gender is an indicator, which is unity for male participants. Education is an indicator, which is unity for participants who at least obtained a college degree. Experience in online labor markets measures the weekly hours a participant spends doing tasks on online labor markets. I standardize the variables for age and experience in online labour markets to facilitate the convergence of the sampler.

treatment dummies. Finally, a third model (M3) controls for socio-economic variables (age, gender, education, and experience in online labor markets).

For each model, the right column of Figure 3 depicts the posterior distributions of the estimated parameters.[27] Estimating only treatment effects (M1) results in credible negative effects of treatments CO and RE on the reserve price to engage in the counting task. The effect of treatment CO has a probability of 1.00 to be negative (MAP $= -0.863$, 0.95-HDI [-1.215, -0.504]) and is credible ($P_{ROPE} = 0$). The effect of treatment RE has a probability of 0.999 to be negative (MAP $= -0.541$, 0.95-HDI [-0.829, -0.171]) and is credible ($P_{ROPE} = 0.007$). The effect of treatment FI, however, has only a probability of 0.851 to be negative (MAP $= -0.285$, 0.95-HDI [-0.656, 0.067]) and too much of the probability mass resides within ROPE ($P_{ROPE} = 0.132$). Moreover, treatment comparisons indicate a negative difference between treatment CO and treatment FI (MAP $= -0.569$, 0.95-HDI [-0.916, -0.2056]), albeit non-credible ($P_{ROPE} = 0.770$). There is no apparent difference between treatment RE and FI (MAP $= -0.191$, 0.95-HDI [-0.559, 0.112]).

The results get stronger when accounting for perceived agreement and its interactions with the treatments (M2). Now all treatment effects are credibly negative. The posterior parameter distributions of treatment CO (MAP $= -1.191$, 0.95-HDI [-1.969, -0.437], $P_{ROPE} = 0.002$), treatment FI (MAP $= -1.075$, 0.95-HDI [-1.840, -0.195], $P_{ROPE} = 0.010$), and treatment RE (MAP $= -0.992$, 0.95-HDI [-1.528, -0.409], $P_{ROPE} = 0.001$) distinctly reside in the negative domain outside ROPE. There is no apparent difference between treatments CO, FI, and RE. Moreover, the indicator for perceived agreement has no clear direction of effect. Its posterior parameter distribution has a probability of 0.555 to be negative (MAP $= -0.052$, 0.95-HDI [-0.625, 0.514]). Interaction effects between perceived agreement and treatment CO (MAP $= 0.363$, 0.95-HDI [-0.428, 1.307], $P_{ROPE} = 0.106$) and between perceived agreement and treatment FI (MAP $= 0.849$, 0.95-HDI [-0.042, 1.790], $P_{ROPE} = 0.030$) also do not exist. By contrast, the posterior parameter distribution of the interaction between perceived agreement and treatment RE has a probability of 0.997 to be positive (MAP $= 1.055$, 0.95-HDI [0.308, 1.711]) and is credible ($P_{ROPE} = 0.006$). As 69.62% (165/237) of agents, who had a WTB choice, perceived the communication with the principal as agreement, this credible interaction effect of perceived agreement and treatment RE explains why the aggregate analysis did not find an effect for treatment RE.

The results of M2 are robust to adding controls for age, gender, education, and experience in online labor markets to the model (M3), as illustrated by the considerable overlap between the 0.95-HDIs of M2 and M3 in Figure 3.

> **Result:** When appropriately accounting for agents' perception about agreements, compensation and fines reduce agents' reserve price for acting opportunistically. That is, WTB increases under compensation and fines. Recovery procedures, however, condition on perceived agreements such that WTB increases in its absence and decreases in its presence.

[27] All three simulations converge very well to the posterior distribution. The largest potential scale reduction factor among the simulated parameters of all three models is close to one, $\max(\hat{R}) < 1.001$. The smallest ratio of the effective sample size to the number of draws is $\min(n_{eff}/N) = 0.938$.

### 4.3.3 Principals' punishment decisions

**Table 3:** Punishment rates by treatment.

| | all | | agreement = 1 | | agreement = 0 | |
|---|---|---|---|---|---|---|
| | agent sends back | | agent sends back | | agent sends back | |
| treatment | 6 tokens | 2 tokens | 6 tokens | 2 tokens | 6 tokens | 2 tokens |
| BL | 0.347 | 0.551 | 0.331 | 0.525 | 0.500 | 0.800 |
| CO | 0.339 | 0.526 | 0.378 | 0.560 | 0.139 | 0.350 |
| FI | 0.249 | 0.521 | 0.255 | 0.526 | 0.100 | 0.400 |
| RE | 0.419 | 0.518 | 0.601 | 0.703 | 0.286 | 0.384 |

Table 3 reports the mean punishment rates by treatment and conditional on whether the principal perceives the communication with the agent as agreement and conditional on the agent's action. Across all treatments and independent of whether the principal evaluates the communication with the agent as agreement, mean punishment rates for non-cooperative play are higher than for cooperative play. Moreover, in all treatments other than BL, mean punishment rates are higher when the principal perceives an agreement with the agent.

In the present context, I am most interested in the effects of the different remedies on the punishment rates for non-cooperative play. To quantify these effects, I estimate three Tobit models. Model M1 estimates the effects of CO, FI, and RE treatment indicators on the punishment rate given non-cooperative play. Model M2 adds an indicator variable that takes unity if the principal perceives the communication with the agent as agreement and interaction effects between perceived agreement and the treatment dummies. Model M3 additionally controls socio-economic variables.

The left column of Figure 4 reports the corresponding posterior parameter distributions.[28] As Figure 4 shows, estimating only treatment effects (M1) yields no existing effects. The posterior probability mass of the parameter distributions covers a substantial area in both the positive and negative domains. However, a clear pattern emerges when considering perceived agreement and its interactions with the treatments.

When agents do not perceive an agreement to exchange tokens cooperatively, treatments CO and RE have a credible negative effect on the punishment rate for non-cooperative choices. The parameter posterior of treatment CO has a probability of 0.990 to be negative (MAP $= -1.742$, 0.95-HDI [-3.158, -0.190]) and lies sufficiently outside ROPE ($P_{\text{ROPE}} < 0.003$). The effect of treatment RE has a probability of 0.993 to be negative (MAP $= -1.306$, 0.95-HDI [-2.837, -0.229]), and a sufficient amount of the probability mass resides outside ROPE ($P_{\text{ROPE}} < 0.003$). The evidence does not indicate any direction of effect of treatment FI on punishment rates absent agreement (MAP $= -1.512$, $P(\theta < 0) = 0.948$, 0.95-HDI [-3.855, 0.292]). Perceived agreement alone also has no clear direction of effect on the punishment rate for agents' non-cooperative choices (MAP $= -1.125$, $P(\theta < 0) = 0.968$, 0.95-HDI [-2.464, 0.133]). However, the estimation shows credible positive interaction effects between perceived agreement and treatments CO

[28] The simulations converge very well to the posterior distribution. The largest potential scale reduction factor among the simulated parameters of all three models is close to one, $\max(\hat{R}) < 1.001$. The smallest ratio of the effective sample size to the number of draws is $\min(n_{\text{eff}}/N) = 0.942$.

and RE, respectively. The interaction effect of perceived agreement and treatment CO has a probability of 0.991 to be positive (MAP $= 1.668$, 0.95-HDI [0.217, 3.342]) and lies sufficiently outside ROPE ($P_{ROPE} < 0.003$). Similarly, the interaction effect of perceived agreement and treatment RE has a probability of 0.998 to be positive (MAP $= 1.868$, 0.95-HDI [0.488, 3.292]) and lies sufficiently outside ROPE ($P_{ROPE} < 0.002$). By contrast, the evidence does not show an interaction effect between perceived agreement and treatment FI (MAP $= 1.444$, $P(\theta > 0) =$ 0.946, 0.95-HDI [-0.342, 3.912]). As indicated by the substantial overlap of posterior parameter distributions of M2 and M3, these results are robust to adding controls in model M3.[29]

> **Result:** Compensation and recovery regimes reduce punishment rates absent perceived agreement but increase punishment when the principal perceives the communication with the agent as agreement. The results do not suggest convincing evidence for any effect of fines on punishment rates with or without perceived agreement.

### 4.4 Post-experiment survey responses

Finally, I look at participants' reported agreement with the three statements about the moral blameworthiness of participating in the counting task. Figure 5 depicts the mean reported importance rating by treatment and the 95%-confidence-interval for each statement. The visual results suggest that treatment RE reduces the moral concern associated with disappointing the other's expectations when participating in the counting task. Relative to treatment FI, treatments CO and RE may reduce the moral concern associated with going back on one's word when participating in the counting task. Moreover, treatment RE also appears to reduce the importance of positive reciprocity for the moral blameworthiness of participating in the counting task.

I quantify the effects by estimating Bayesian Tobit models. For the reported agreement with each moral concern about participating in the counting task, I employ the same model structures as before. That is, I estimate (1) treatment effects only, (2) I add perceived agreement and its interaction with the treatments, and (3) I add control variables in a third estimation.

Regarding the concern for disappointing expectations, treatment RE has a credible negative effect on the reported agreement in all three estimations. In the estimation with interaction effect and control variables, for instance, the posterior parameter has a probability of 0.999 to be negative (MAP $= -32.883$, 0.95-HDI [-52.272, -11.804]) and is credible ($P_{ROPE} < 0.002$). Relative to the control condition, no other credible treatment effects occur. While the difference between the posterior parameter values of treatment RE and FI is credibly negative (MAP $= -22.565$, 0.95-HDI [-38.458, -16.981], $P_{ROPE} < 0.001$) when estimating treatment effects only, the two more detailed estimations do not deliver consistent evidence.

[29] For completeness, I have conducted the same analysis of punishment rates, given that the agent acts cooperatively. The right column of Figure 4 illustrates the results. Except for the interaction effect between perceived agreement and treatment RE, none of the independent model components has a credible effect on punishment rates. However, the interaction effect between perceived agreement and treatment RE has a probability of 0.985 to be positive (MAP $= 1.124$, 0.95-HDI [0.114, 2.193]) and is credible ($P_{ROPE} < 0.005$).

Regarding the concern for going back on one's word, estimating treatment effects only yields evidence for a credible negative effect of treatment CO on the reported agreement. The posterior parameter distribution has a probability of 0.995 to be negative (MAP $= -13.626$, 0.95-HDI [-24.380, -3.217]) and is credible ($P_{ROPE} = 0.024$). A negative effect of treatment RE relative to the control condition exists (MAP $= -10.099$, 0.95-HDI [-20.168, -0.787]) but is not credible for practical purposes ($P_{ROPE} = 0.055$). The same pattern occurs relative to treatment FI. While the effect of treatment CO is credibly negative (MAP $= -16.584$, 0.95-HDI [-25.685, -4.932], $P_{ROPE} = 0.011$), treatment RE has a negative effect that is (barely) not credible (MAP $= -12.476$, 0.95-HDI [-22.866, -3.255], $P_{ROPE} = 0.027$). Note that estimating the two more elaborate models does not yield any meaningful effects.

Lastly, regarding the concern of finding the Sender deserving of reciprocal generosity, estimating treatment effects yields a credible negative effect of treatment RE on the agreement with the reciprocity concern. The posterior parameter distribution has a probability of 1 to be negative (MAP $= -20.881$, 0.95-HDI [-31.115, -11.684]) and is credible ($P_{ROPE} < 0.001$). Treatment RE also has a credible negative effect relative to treatments CO (MAP $= -15.204$, 0.95-HDI [-25.399, -6.323], $P_{ROPE} = 0.006$) and FI (MAP $= -23.182$, 0.95-HDI [-32.019, -12.471], $P_{ROPE} < 0.001$). When adding perceived agreement and its interaction with the treatments to the estimation, the effect of treatment RE relative to the control condition BL (MAP $= 27.349$, 0.95-HDI [-44.407, -7.375], $P_{ROPE} = 0.006$) and treatment CO (MAP $= -21.855$, 0.95-HDI [-40.590,-2.071], $P_{ROPE} = 0.025$) remains credibly negative. The difference between treatments RE and FI is not credible anymore, however. Note that all effects go away when estimating the full-fledged model, including control variables.

**Figure 2:** Results of Bayesian regression analysis of cumulative agents' WTB shares (left column), cumulative principals' WTB beliefs (middle column), and the aggregate WTB choice-belief gap (right column). Left and middle columns present results from beta models (Ferrari and Cribari-Neto, 2004), whereas the right column present results from linear models. The choice-belief gap is the standardized difference between cumulative principals' WTB beliefs and cumulative agents' WTB shares. Treatment variables are dummies for the respective treatments. Net profits are scaled following Gelman (2008). Each panel depicts posterior parameter distributions, 0.95-HDIs with the maximium a priori probability estimate, and the region of practical equivalence (ROPE). ROPE bounds are set to $\pm 0.181$ for beta models and $\pm 0.1$ of one standard deviation of the dependent variable otherwise. ROPE bounds adjusted for scaling of independent variables.

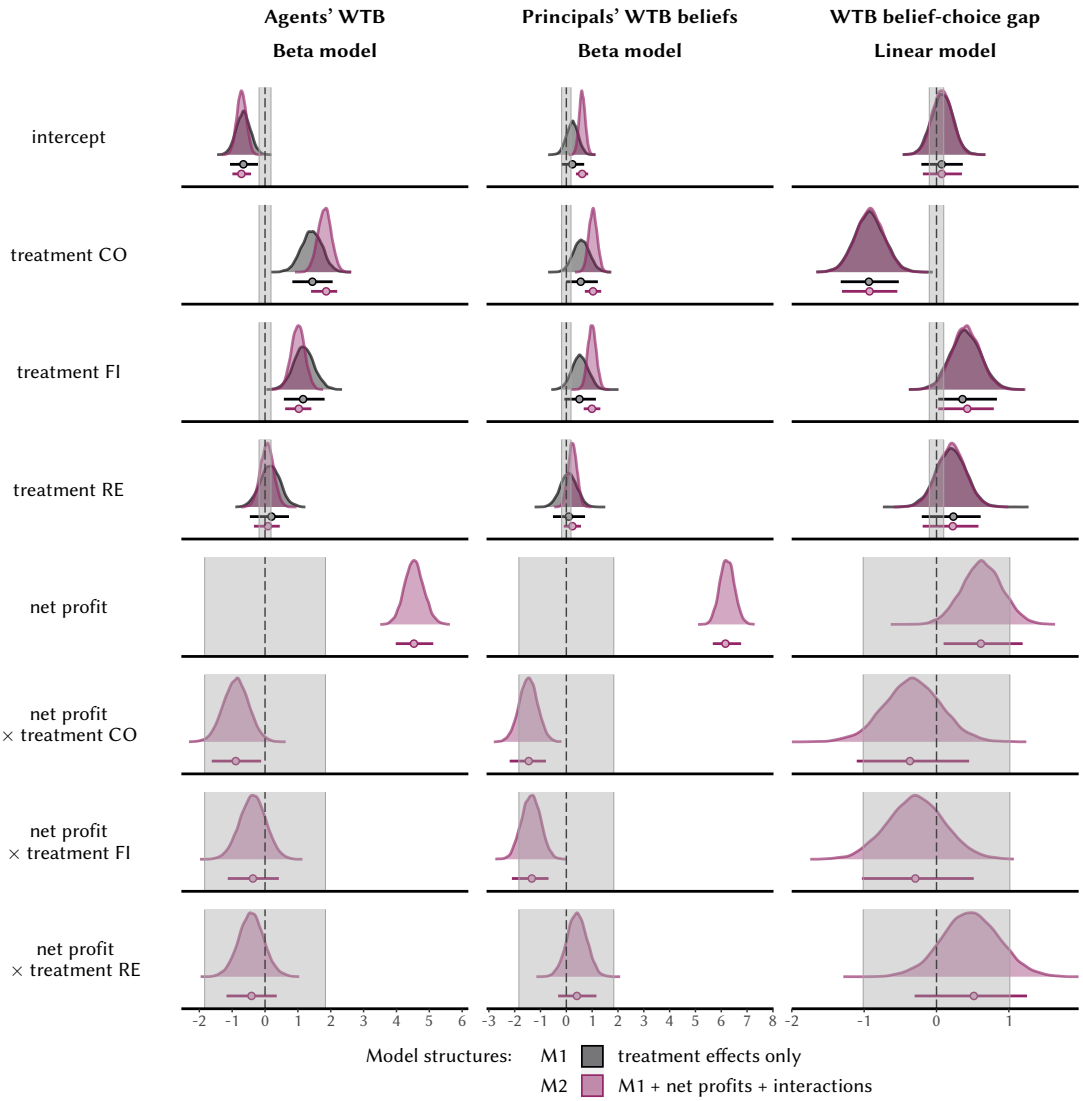


**Figure 3:** Results of Bayesian regression analysis of principal's choice to initiate transactions by sending four tokens to the agent (left column) and agent's minimal required net profits in order to engage in the counting task, i.e., her willingness-to-breach (right column). Both columns show results from a linear model. Agent's reserve price is standardized. Treatment variables are dummies for the respective treatments. The variable "agreement" indicates whether the participant evaluates to communication between the players as agreement. Each panel depicts posterior parameter distributions, 0.95-HDIs with the maximium a priori probability estimate, and the region of practical equivalence (ROPE). ROPE bounds are set to $\pm 0.1$ of one standard deviation of the dependent variable.

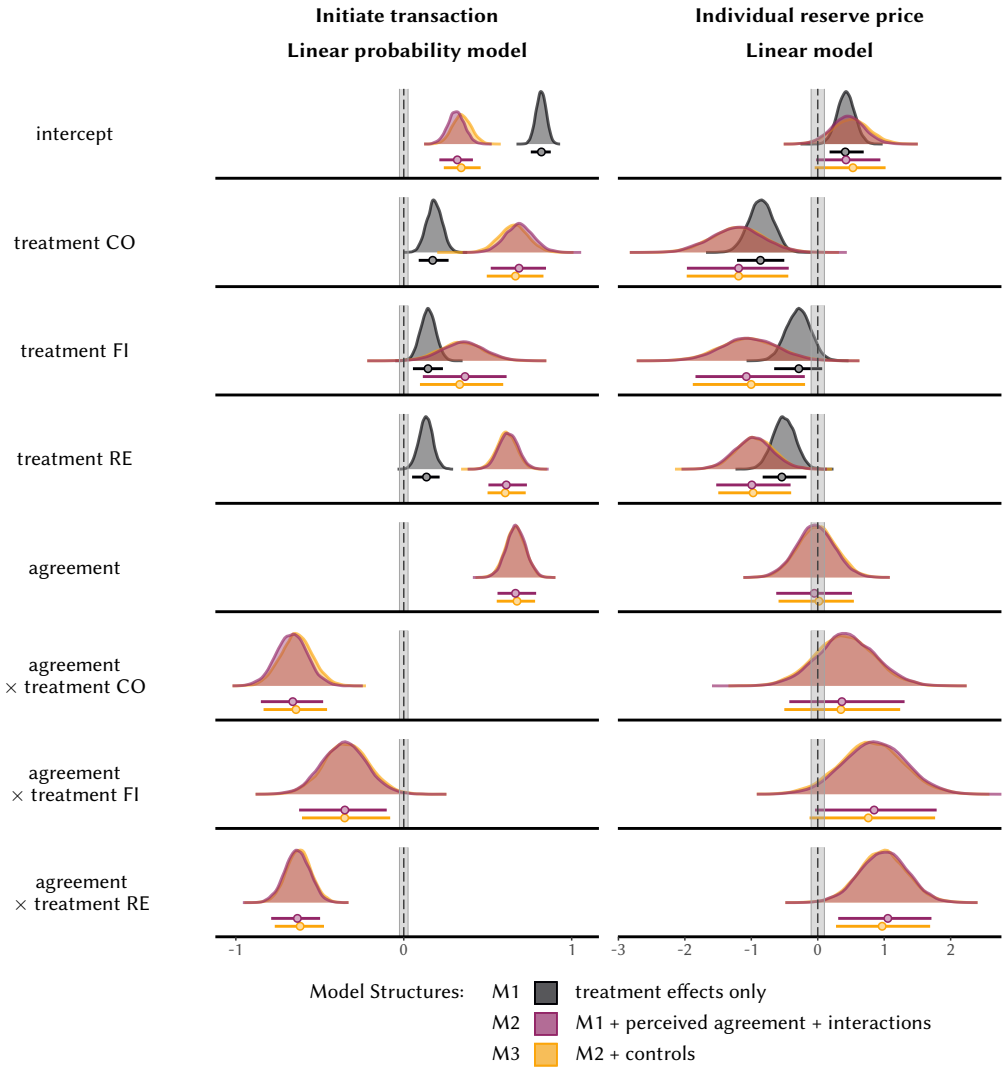


**Figure 4:** Results of Bayesian regression analysis of punishment proportions for non-cooperative decisions (left column) and cooperative decisions (right column). Both columns show results from Tobit models. Treatment variables are dummies for the respective treatments. The variable "agreement" indicates whether the participant evaluates to communication between the players as agreement. Each panel depicts posterior parameter distributions, 0.95-HDIs with the maximium a priori probability estimate, and the region of practical equivalence (ROPE). ROPE bounds are set to $\pm 0.1$ of one standard deviation of the dependent variable.

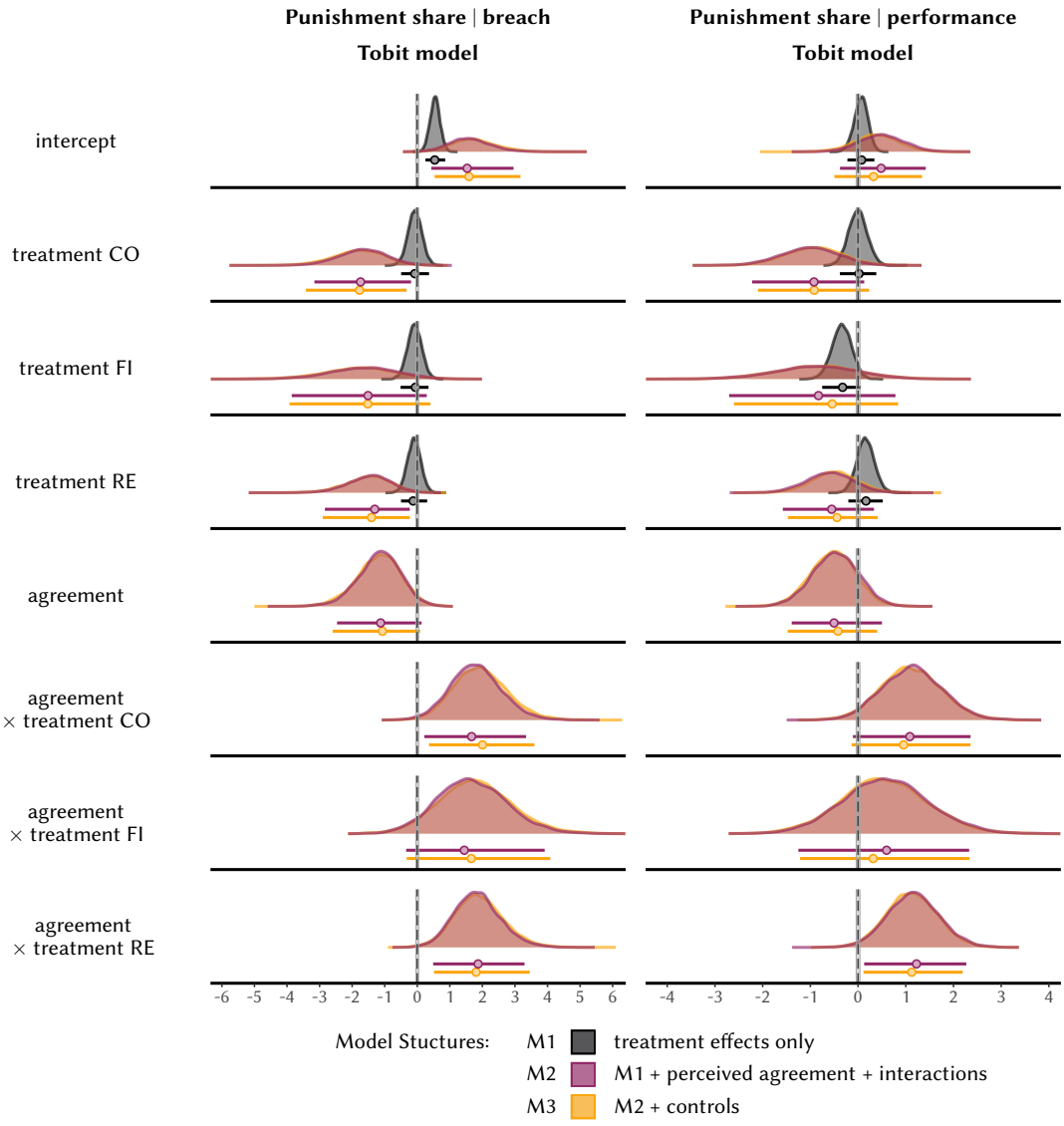
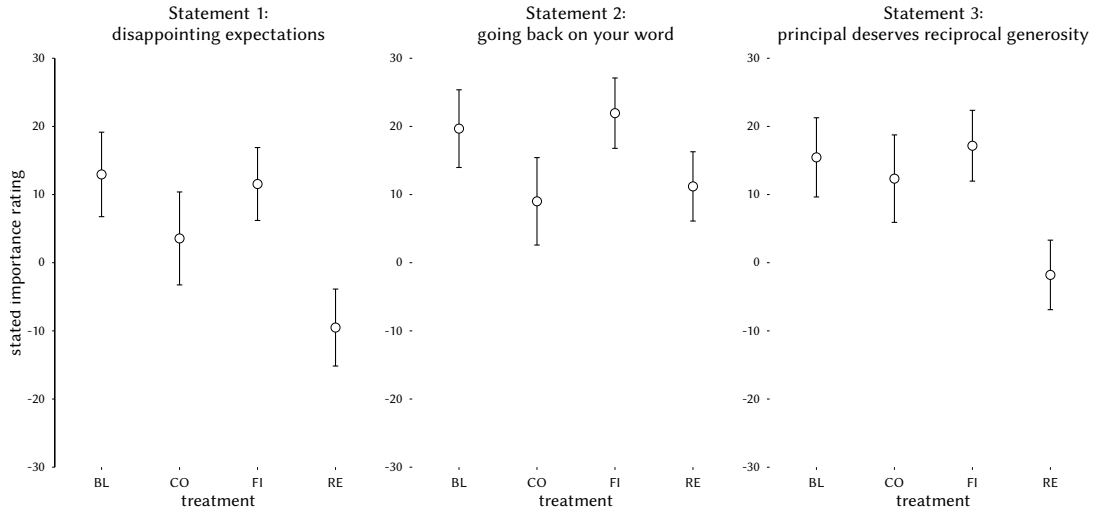


**Figure 5:** Results of the post-treatment questionnaire. Each panel depicts the mean of the reported importance ratings for that statement by treatment and the corresponding 95%-confidence intervals.



# 5 Discussion

The experiment yields multiple vital insights. On the aggregate level, compensations and fines elevate the cumulative share of agents willing to forego the current transaction for a more lucrative one. This insight alone is an important finding as it qualifies earlier research on breach of contract (e.g.: Wilkinson-Ryan, 2015). The recovery mechanism, by contrast, does not yield such an effect. This pattern of results suggests that the element of fines drives the effect of compensation, which is reminiscent of Gneezy and Rustichini (2000). However, the effect of compensation is greater than the effect of fines. That is, one or more additional triggers may explain the increase in willingness-to-breach under the compensation mechanisms. I did not design this experiment to reveal such nuanced triggers because I had no corresponding predictions. With the current insights, however, doing so is an invaluable task for future research.

On the aggregate level, I also find that compensations and fines increase principals' beliefs about the agents' willingness-to-breach. These remedies thus have a behavioral effect on both the agent and the principal. Importantly, principals' beliefs and agents' actual choices coincide only in the compensation treatment, which strongly reduces the belief-choice gap. In aligning beliefs and choices of contract parties, compensation mechanisms thus create a form of institutional debiasing (cf.: Arlen and Tontrup, 2015).

On the individual level, the results provide more insights into the effects of contract remedies on the choices of principal and agent. First, consider the interaction between the remedies and a perceived agreement to exchange tokens cooperatively. Absent perceived agreement, compensation, fine, and recovery mechanisms increase the trusting behavior of principals. This finding is intuitive but essential nevertheless, insofar as compensation and recovery insure the principal against the opportunistic behavior of the agent. Fines, however, yield a similar effect without insuring the principal. Possibly, the principal counts on fines being harsh enough to discipline the agent. Second, perceived agreement to exchange tokens cooperatively also elevates the trusting behavior of the principal. However, any remedies governing opportunistic behavior mute a perceived agreement's effect. I interpret this finding as evidence that contract remedies and agreements are substitutes.

Second, the results on the individual level provide strong support for Hypotheses 1 and 3. Fines and compensations cause a decrease in agents' reserve prices for participating in the counting task instead of finishing the transaction with the principal. In other words, compared to the baseline treatment, agents require less profit to switch to the more lucrative alternative. Note that this is independent of perceived agreement. By contrast, the effect of the recovery procedure conditions on a perceived agreement. When agents did not perceive an agreement with the principal to exchange tokens cooperatively, the recovery procedure reduced the reserve price. Conversely, when agents perceive such an agreement, the recovery procedure increases the reserve price. This result contrasts Hypothesis 2 to the extent that promise-induced guilt aversion inspires this prediction. When the agent believes in agreeing with the principal to exchange tokens cooperatively, the recovery mechanisms should reduce guilt and lead to a lower reserve price. This result is even more noteworthy as I, in contrast to a robust prior research stream (e.g.: Charness and Dufwenberg, 2006; Vanberg, 2008; Ederer and Stremitzer, 2017), observe no positive effect of a perceived agreement on reserve prices, even without any

remedy. I cautiously conjecture that an increased social distance between the participants in this experiment may drive this finding. Opposite to the prior literature, I conducted the experiment on AMT instead of a computer laboratory. Whether this difference increases social distance and whether social distance indeed affects promise-induced guilt aversion is another open venue for future research. Nevertheless, when some remedial mechanism would burden the agents with switching costs, offset the principal's loss, or both – in particular absent perceived agreement – agents exhibit a higher tendency to opt for the opportunistic choice. While not all agents switched to the more lucrative transaction for the tiniest profit, the simple remedies tested in the experiment still encouraged efficient breach.

Third, the data support Hypothesis 4. Compensation and recovery mechanisms reduce the principals' punishment rate for opportunistic choices, absent perceived agreement. Under these conditions, the principal apparently evaluates opportunistic actions as less morally reprehensible. I interpret this result as supporting evidence for compensatory damages' indemnification and vindication function. The reduction of punishment rates seems to be tied to the recovery element of compensatory damages because the fines do not exhibit a similar effect. Interestingly, I pick up positive interaction effects between perceived agreement and these remedies, i.e., compensation and recovery. If the principal thought she had an agreement with the agent, she would punish the agent more under compensation and recovery mechanisms than without any remedy. I predicted such an effect for the other combination of remedies, namely compensation and fines, because the remedies' negative consequences for the agent emphasize the agent's culpability (Hypothesis 5). Given perceived agreement, the increased punishment rates for opportunistic choices when compensation and recovery procedures insure the principal suggest a different interpretation, however. Suppose compensation and recovery communicate a social norm that permits otherwise prohibited behavior because this behavior creates less or no harm. Such a norm would facilitate breach and simultaneously render breach less reprehensible. Note that this norm is notably different from a norm that permits otherwise prohibited behavior in exchange for a price, such as fines (Gneezy and Rustichini, 2000; Kornhauser et al., 2020). By expressly agreeing to the cooperative exchange, both parties negotiate away from that social norm and shift the reference point for the principal's expectations. As a consequence, breaking the agreement now hurts the principal even more than when the parties had not mutually overridden the social norm permitting breach. Whether agreement to an action that runs contrary to what an underlying social norm suggests in fact creates an adverse response when one of the parties breaks that agreement deserves intensified study. As an implication, compensatory damages may better fulfill their indemnification and vindication function when parties transact with standard form contracts, and the agreement carries less meaning *inter partes*. By contrast, when parties expressly negotiate highly specific contracts and agree to individual terms, principals appear to find breach more reproachable.

## 6 Conclusion

In this paper I report the results of a controlled experiment that provides an institution test on how compensatory damages influence participants' propensity to forego a lucrative transaction for increased profit. While efficient breach theory recommends expectation damages as remedy

for breach of contract, a broad prior literature shows that people tend to keep their promises, honor their agreements, and see through transactions, although doing so is individually costly. To the extent that people are hesitant to act opportunistically, these findings draw into question the heuristic value of efficient breach theory.

In the experiment, I elicit the reserve price of second movers (agents) in a trust game for foregoing the cooperative solution to the game. Second movers receive a surprising opportunity to play a possibly more lucrative game rather than finishing the transaction with the first movers (principals). The experiment also features pre-play communication between principal and agent. Moreover, principals can punish agents for cooperative or non-cooperative choices.

As an important contribution to the prior literature, my results indicate that compensatory damages, fines, and recovery procedures increase participants' willingness-to-breach, i.e., they reduce agents' reserve prices for acting opportunistically. The data also suggest that compensatory damages align principals' beliefs about opportunistic decisions with the actual decisions of agents. Finally, the experiment reveals that the recovery element of compensatory damages has a differentiated effect on how morally reprehensible principals evaluate opportunistic actions of agents. Absent agreement, principals punish opportunistic agents less when they are insured by compensation or recovery regimes. However, although being insured principals punish opportunistic agents more when the parties expressly agreed upon mutually-beneficial exchange.

Insofar as the mechanisms exposed to this institution test increase agents' willingness-to-breach, the results speak to the fundamental assumption of efficient breach theory that contracting parties are poised to breach for profit. However, the results presented here do not only preserve the heuristic value of efficient breach. Rather, they suggest an important twist. Expectation damages as recommended by efficient breach theory provide efficient incentives, but more importantly this remedy also facilitates breach in the first place. Insofar as compensatory damages align beliefs of principals about reserve prices with actual agents' reserve prices, the results speak to another form institutional debiasing, which has been an important topic in the recent literature as well. Finally, insofar as remedies reduce the principals' punishment of opportunistic behavior, the results emphasizes the indemnification and vindication function of compensatory damages.

The experiment presented here is designed as testbed for contract remedies. Therefore I cannot disentangle potentially relevant behavioral channels. With this limitation in mind, I make multiple suggestions for future research throughout.

# References


Ariely, D., Bracha, A., and Meier, S. (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, 99(1):544–555.

Arlen, J. and Tontrup, S. (2015). Does the Endowment Effect Justify Legal Intervention? The Debiasing Effect of Institutions. *Journal of Legal Studies*, 44(1):143–182.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in Experiments: A Theoretical Analysis. *Journal of Political Economy*, 126(4):1472–1503.

Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). Self-fulfilling Property of Trust: An Experimental Study. *Theory and Decision*, 63:349–388.

Balafoutas, L. and Sutter, M. (2017). On the Nature of Guilt Aversion: Insights from a New Methodology in the Dictator Game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Balliet, D. (2010). Communication and Cooperation in Social Dilemmas: A Meta-Analytic Review. *Journal of Conflict Resolution*, 54(1):39–57.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, 97(2):170–176.

Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2013). Shaping Beliefs in Experimental Markets for Expert Services: Guilt Aversion and the Impact of Promises and Money-burning Options. *Games and Economic Behavior*, 81:145–164.

Becker, G. M., Degroot, M. H., and Marschak, J. (1964). Measuring Utility by a Single-response Sequential Method. *Behavioral Science*, 9(4):226–232.

Bellemare, C., Sebald, A., and Suetens, S. (2017). A Note on Testing Guilt Aversion. *Games and Economic Behavior*, 102:233–239.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge.

Bigoni, M., Bortolotti, S., Parisi, F., and Porat, A. (2017). Unbundling Efficient Breach: An Experiment. *Journal of Empirical Legal Studies*, 14(3):527–547.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1):166–193.

Brandts, J. and Charness, G. (2011). The Strategy Versus the Direct-response Method: a First Survey of Experimental Comparisons. *Experimental Economics*, 14:375–398.

Brekke, K. A., Kverndokk, S., and Nyborg, K. (2013). An Economic Model of Moral Motivation. *Journal of Public Economics*, 87(9):1967–1983.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science*, 351(6280):1433–1436.

Charness, G. and Dufwenberg, M. (2006). Promises and Partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Dufwenberg, M. (2010). Bare Promises: An Experiment. *Economic Letters*, 107(2):281–283.

Charness, G. and Dufwenberg, M. (2011). Participation. *American Economic Review*, 101(4):1211–1237.

Charness, G. and Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *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.

Chen, Y., Kartik, N., and Sobel, J. (2008). Selecting Cheap-Talk Equilibria. *Econometrica*, 76(1):117–136.

Cialdini, R. B. (2007). *Influence: The Psychology of Persuasion*. Harper Collins, New York, 3rd edition.

Colling, L. J. and Szűcs, D. (2018). Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology*.

Cooter, R. and Ulen, T. S. (2016). *Law & Economics*. Berkeley Law Books, Berkeley, 6th edition.

Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A Tractable Model of Reciprocity and Fairness. *Games and Economic Behavior*, 59:17–45.

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71(9):1–25.

Depoorter, B. and Tontrup, S. (2012). How Law Frames Moral Intuitions: The Expressive Effect of Specific Performance. *Arizona Law Review*, 54:673–717.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, 6(3):274–290.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The Framing of Games and the Psychology of Play. *Games and Economics Behavior*, 73:459–478.

Ederer, F. and Stremitzer, A. (2017). Promises and Expectations. *Games and Economic Behavior*, 106:161–178.

Eisenberg, M. A. (2018). *Foundational Principles of Contract Law*. Oxford University Press, New York.

Ellingsen, T. and Johannesson, M. (2004). Promises, Threats and Fairness. *The Economic Journal*, 114(495):397–420.

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

Fehr, E. and List, J. A. (2004). The Hidden Costs and Returns of Incentives – Trust and Trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5):743–771.

Fehr, E. and Rockenbach, B. (2003). Detrimental Effects of Sanctions on Human Altruism. *Nature*, 422:137–140.

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

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press, Stanford.

Gelman, A. (2008). Scaling Regression Inputs by Dividing by Two Standard Deviations. *Statistics in Medicine*, 27(15):2865–2873.

Gneezy, U. and Rustichini, A. (2000). A Fine is a Price. *The Journal of Legal Studies*, 29(1):1–17.

Harbring, C. (2006). The Effect of Communication in Incentive Systems–An Experimental Study. *Managerial and Decision Economics*, 27(5):333–353.

Hauge, K. E. (2016). Generosity and Guilt: The Role of Beliefs and Moral Standards of Others. *Journal of Economic Psychology*, 54:35–43.

Hoeppner, S., Freund, L., and Depoorter, B. (2017). The Moral-Hazard Effect of Liquidated Damages: An Experiment on Contract Remedies. *Journal of Institutional and Theoretical Economics*, 173(1):84–105.

Hoppe, E. I. and Schmitz, P. W. (2018). Hidden Action and Outcome Contractibility: An Experimental Test of Moral Hazard Theory. *Games and Economic Behavior*, 109:544–564.

Ioannidis, J. P. A., Stanley, T. D., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, 127(605):F236–F265.

Ismayilov, H. and Potters, J. (2012). Promises as Commitments. *Tilburg University, Center for Economic Research, Discussion Paper*, 2012-064.

Isoni, A., Loomes, G., and Sugden, R. (2011). The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Comment. *American Economic Review*, 101(2):991–1011.

Kartik, N. (2009). Strategic Communication with Lying Costs. *Review of Economic Studies*, 76:1359–1395.

Kessler, J. B. and Leider, S. (2012). Norms and Contracting. *Management Science*, 58(1):62–77.

Klass, G. (2014). Efficient Breach. In Klass, G., Letsas, G., and Saprai, P., editors, *The Philosophical Foundations of Contract Law*, pages 362–387. Oxford University Press, Oxford.

Kornhauser, L., Lu, Y., and Tontrup, S. (2020). Testing a Fine is a Price in the Lab. *International Review of Law and Economics*, 63. DOI: 10.1016/j.irle.2020.105931.

Kruschke, J. K. (2010). Bayesian Data Analysis. *WIREs Cognitive Science*, 1(5):658–676.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis*. Academic Press, Boston, 2 edition.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

Levine, D. K. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, 1(3):593–622.

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., and Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10:2767.

Malmendier, U., te Velde, V. L., and Weber, R. A. (2014). Rethinking Reciprocity. *Annual Review of Economics*, 6(1):849–874.

Mellström, C. and Johannesson, M. (2008). Crowding Out in Blood Donation: Was Titmuss Right? *Journal of the European Economic Association*, 6(4):845–863.

Metcalf, C., Satterthwaite, E. A., Dillbary, J. S., and Stoddard, B. (2020). Is a Fine Still a Price? Replication as Robustness in Empirical Legal Studies. *International Review of Law and Economics*, 63. DOI: 10.1016/j.irle.2020.105906.

Mischkowski, D., Stone, R., and Stremitzer, A. (2019). Promises, Expectations, and Social Cooperation. *Journal of Law and Economics*, 62(4):687–712.

Open Science Collaboration (2012). An Open, Large-scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspective of Psychological Science*, 7:528–530.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science*, 349(6251):aac4716.

Plott, C. R. and Zeiler, K. (2005). The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations. *American Economic Review*, 95(3):530–545.

Plott, C. R. and Zeiler, K. (2011). The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Reply. *American Economic Review*, 101(2):1012–1028.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Technical University Vienna, Vienna.

Predmore, C., Topyan, K., and Apadula, L. T. (2021). Impact of Process Misconsception in Becker-DeGroot-Marschak Single Response Value Elicitation Procedures: An Experimental Investigation in Consumer Behavior Using the IKEA Effect. *Economies*, 9(4):173.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83(5):1281–1302.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In Sauermann, H., editor, *Beiträge zur experimentellen Wirtschaftsforschung*, page 136–168. Mohr, Tübingen.

Sutter, M. (2009). Deception Through Telling the Truth?! Experimental Evidence from Individuals and Teams. *The Economic Journal*, 119(534):47–60.

Ulen, T. S. (1984). The Efficiency of Specific Performance: Toward a Unified Theory of Contract Remedies. *Michigan Law Review*, 83(2):341–403.

Vanberg, C. (2008). Why Do People Keep Their Promises? An Experimental Test of Two Explanations. *Econometrica*, 76(6):1467–1480.

Wagenmakers, E., Wetzels, R., Borsboom, D., and van der Maas, H. (2011). Why Psychologists Must Change the Way They Analyze Their Data:The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3):426–432.

Wilkinson-Ryan, T. (2015). Incentives to Breach. *American Law and Economics Review*, 17(1):290–311.