Dear Workshop Participants,

This paper reports results of a vignette study in which subjects tell us how they believe they would respond when faced with various (hypothetical) scenarios along with a closely related experiment in which subjects engage in real interactions and faced monetary consequences for their (and others') decisions. The latter experiment is a pilot study on a subject pool of MTurk workers, which we conducted just last week. We wanted to make sure we were able to replicate the results of the vignette in a setting with real interactions and monetary incentives before conducting a more expensive, full-scale version of the second experiment on student participants in the lab. We will be able to make changes to the design prior to running the full version of that experiment. Thus, your feedback on all aspects of the paper, especially the design of the second experiment, will be particularly valuable to us going forward.

Rebecca

Norm-Based Enforcement of Promises*

Nathan Atkinson[†] ETH Zurich

Rebecca Stone[‡] UCLA

Alexander Stremitzer[§] ETH Zurich

January 22, 2020

Abstract

Considerable evidence suggests that people are internally motivated to keep their promises. But it is unclear whether promises alone create a meaningful level of commitment in many economically relevant situations where the stakes are high. We experimentally study the behavior of third-party observers of an interaction between a potential promisor and potential promisee who have the opportunity to punish, at a cost to themselves, non-cooperative behavior by the potential promisor. Our results suggest that third parties' motivations to punish promise breakers have the same structure as the moral motivations of those deciding whether or not to keep their promises. That is, the same moral reasons that seem to motivate promisors to keep their promises make third-party observers more likely to punish promise breaking. This suggests that underlying promissory norms drive both promise-keeping behavior in the absence of second- and third-party enforcement mechanisms and non-legal enforcement mechanisms that arise when third parties can punish potential promisors in a decentralized fashion. This makes it more likely that promissory norms support cooperative behavior even when the stakes are high.

Keywords: promises, norms, first-party enforcement, second-party enforcement, altruistic punishment.

JEL-Classification: K12, L14, D86, D91, C91.

*The authors are grateful to Gary Charness and William Hubbard for valuable comments. We are also grateful to seminar audiences at Alabama, ETH Zurich, University of Zurich, and the Annual Conference of the Polish Law and Economics Society. We thank Henry Kim for excellent research assistance.

[†]ETH Zurich, Center for Law & Economics, IFW E49, Haldeneggsteig 4, 8092 Zurich, Switzerland, nate.atkinson@gess.ethz.ch.

[‡]UCLA Law School, 385 Charles E. Young Drive, 1242 Law Building, Los Angeles, CA 90095, rebecca.stone@law.ucla.edu (corresponding author).

[§]ETH Zurich, Center for Law & Economics, IFW E49, Haldeneggsteig 4, 8092 Zurich, Switzerland, astremitzer@ethz.ch.

1 Introduction

A series of experimental studies find that in the absence of legal enforcement and reputational concerns promisors are nonetheless motivated to keep their promises even when they have self-interested reasons to break them (Ellingsen and Johannesson 2004, Charness and Dufwenberg, 2006; Vanberg, 2008; Charness and Dufwenberg, 2010). Promisors are also more likely to keep their promises when promisees expect them to keep them or invested more in reliance on their promises (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Ederer and Stremitzer, 2017; Stone and Stremitzer, forthcoming). And potential promisors are more sensitive to others' expectations of cooperation and reliance investments when they promised those others they would cooperate with them than than when they made no such promises (Mischkowski, Stone, and Stremitzer, forthcoming).

In these ways, promisors seem to be motivated by moral reasons when deciding whether or not to keep their promises. They seem to care about keeping promises for their own sake. They care about not thwarting others' expectations of their behavior, even those not induced by their promises, especially when those others may have changed their behavior as a result of such expectations. And they seem to believe that it is particularly important not to thwart such expectations when those expectations were induced by their promises.

But such moral motivations may not guarantee performance of a promise—even one that has been relied upon extensively by the promisee—when, as will often be the case in commercial settings, the promisor has significant self-interested incentives to break it. And, of course, they will have no effect on the behavior of sophisticated self-interested actors who aren't motivated by moral reasons to keep their promises in the first place. If so, then effective second- and third-party enforcement mechanisms will be needed to give such promisors incentives to perform their promises in many economically relevant situations.

It does not, however, follow that promissory norms will therefore be irrelevant to the behavior of such promisors. Centralized enforcement systems like the legal system can be, and arguably are and/or should be, designed to reflect or at least support underlying moral

practices of their subjects.[1]

Decentralized mechanisms of enforcement by disinterested private observers may also be shaped by promissory norms. If the behavior of promisors is driven by underlying moral norms that deem it wrong to break one's promises and more seriously wrong the more that promisees have relied on them, then we should also expect third-party observers to judge a promisor more harshly when she breaks a promise to cooperate than when she simply fails to cooperate with another without having promised to do so, and that they will judge her more harshly for breaking such a promise the more the promisee has relied upon it. In the light of theory and evidence that suggests that persons are willing to altruistically enforce widely held norms, we should then expect that those third parties will be willing to punish the promisor for breaking her promise, even if such punishment comes at some cost to themselves, and to punish the promisor more the greater was the promisee's reliance on the promise.[2] Reputational concerns that track agents' moral beliefs would then make even self-interested promisors act as if they cared about the underlying moral norms that govern promising.

We test this hypothesis using a vignette study and an experiment with real interactions and monetary incentives. Both experiments yield results that are in large part in line with our hypothesis. Subjects show a greater willingness to punish non-cooperative behavior, even at some cost to themselves, when the non-cooperative agent broke a promise to cooperate than when she made no such promise. In line with the terminology set out in Mischkowski, Stone, and Stremitzer (forthcoming), we call this the promising per se effect. Subjects also show a greater willingness to punish non-cooperative behavior the more the other party invested in anticipation of cooperation regardless of whether the non-cooperative agent made a promise. We call this the reliance per se effect. Finally, the effect of greater reliance on the third-party's willingness to punish non-cooperative behavior is enhanced when it breaks a promise

[1] See, for instance Restatement 2d of Contracts, Section 1 (1981) (defining a contract as an enforceable promise or set of promises).

[2] Gintis et al. (2005) provide an overview of the theory and evidence that supports the existence of dispositions towards strong reciprocity. Fehr and Fischbacher (2004) provide experimental evidence of subjects' propensities to punish third parties at a cost to themselves.

to cooperate. We call this the interaction effect. Thus, our results suggest that promising and reliance shape a third party's willingness to punish promise breaking in much the same way that previous experimental evidence suggests that they shape a potential promisor's internal willingness to keep her promises in the first place.

Taking these results seriously has an important implication. They suggest that underlying promissory norms influence not only the internal motivations of promisors (first-party enforcement mechanisms, as they are sometimes referred to) but also decentralized third-party mechanisms of enforcement. This implication is in line with Hart & Moore's (2008) argument that contracts set reference points or expectations of contracting parties such that disappointed promisees react by punishing the promisor when these expectations are confounded by "shading" on their performance. Hart and Moore's (2008) argument concerns second-party enforcement—that is, enforcement by the party who has been wronged by a breach of promise. Our results suggest third-party enforcement behavior may similarly be shaped by underlying promissory norms.

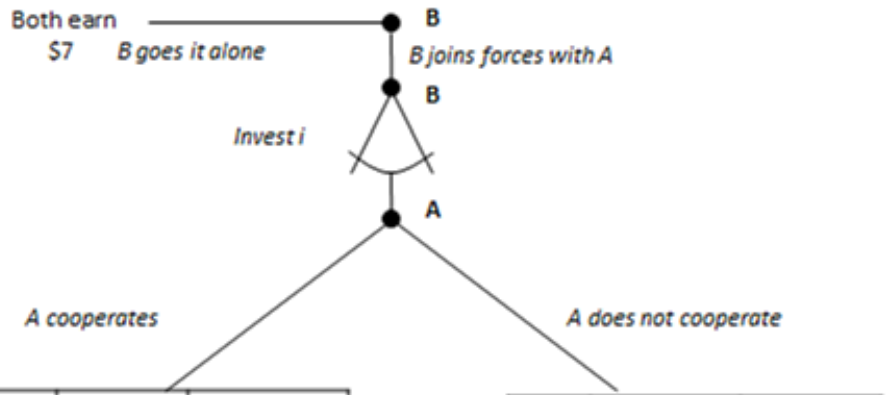
Our results are also in line with the growing body of theoretical and empirical results that suggest that altruistically enforced norms are central determinants of cooperative behavior (e.g. Gintis et al., 2005; Fehr and Fischbacher, 2004).

The remainder of the paper is organized as follows. Section 2 describes the game that forms the bases of both experiments and our hypotheses. Section 3 describes the design of each experiment and its associated procedures. Section 4 describes our results. Section 5 provides further discussion of those results. Section 6 concludes.

2 Trust Game with Punishment and Hypotheses

Both our experiments are built on the trust game that is depicted in Figure 1. In this game, Player B must decide whether to join forces with Player A in a cooperative venture or pursue an alternative project alone. Joining forces with A is risky for B because A might exploit B and take the profits from the venture for herself. But if A decides not to exploit B, both do better than if B had pursued the alternative project alone. If B decides to join forces

with A, cooperation increases B's payoff but reduces A's, and aggregate payoffs are highest when A cooperates. B's payoff also depends on B's investment. Higher investment initially increases and then decreases B's payoff if A cooperates and it always decreases B's payoff, and does so at a faster rate, if A doesn't cooperate.



i	A's profits	B's profits
0	12	12
1	12	12.25
2	12	12.00
3	12	11.75
4	12	11.50
5	12	11.25
6	12	11.00

i	A's profits	B's profits
0	15	6
1	15	5
2	15	4
3	15	3
4	15	2
5	15	1
6	15	0

Figure 1: Game Witnessed by Third Party

We add to the basic trust game an initial communication stage and a final punishment stage. The communication stage occurs before B makes his initial decision whether to join forces with A and gives A the opportunity to promise B that she will cooperate with him if B joins forces with her. The punishment stage occurs after A and B play the trust game and involves a disinterested third party—Player C—who has observed the interaction between A and B. During this final stage, C has the opportunity to punish A by reducing A's payoff, at some cost to himself.

We assume that C's punishment preferences are influenced by the same underlying promissory norms that previous experiments suggest drive persons to keep their promises absent a self-interested reason to do so. Thus, just as people seem to be more willing to

cooperate when they promised to do so all else equal, we hypothesize that third parties will be more willing to punish the non-cooperative behavior of one who promised to cooperate than one who didn't make such a promise (the promise per se effect). Just as people seem to be more willing to cooperate the greater are recipients' expectations of cooperation and ensuing reliance investments, even when they made no promises to cooperate, we hypothesize that all else equal third parties will be more willing to punish non-cooperation as investment levels of recipients increase (the reliance per se effect). Finally, just as people seem to be even more influenced by greater expectations of cooperation and reliance by recipients when they promised recipients they would cooperate, we hypothesize that the effect of greater reliance on the willingness of third parties to punish non-cooperation will be greater when that non-cooperation breaks a promise to cooperate than when no such promise was made (the interaction effect).[3] Hypothesis 1 summarizes.

Hypothesis 1 *C's willingness to punish non-cooperation will be higher when A made a promise than when A made no promise for all levels of B's investment (H1.1). C's willingness to punish non-cooperation will increase with B's investment level whether or not A made a promise. (H1.2) This increase will be greater if A made a promise than if A made no promise (H1.3).*

Given that Hypothesis 1 arises from our conjecture that C's willingness to punish non-cooperation is driven by shared norms, we also hypothesize that subjects' views about the appropriateness of A's actions will exhibit a parallel structure. Hypothesis 2 summarizes.

Hypothesis 2 *Perceptions of the inappropriateness of A's non-cooperation will be higher when A made a promise than when A made no promise for all levels of B's investment (H2.1). Perceptions of the inappropriateness of A's non-cooperation will be increasing in the level of B's investment whether or not A made a promise (H2.2). This increase will be greater if A made a promise than if A made no promise (H2.3).*

[3]Our hypotheses would follow from a model of the observer's preferences that resembles in all important respects the model of promisor's preferences in Mischkowski, Stone, and Stremitzer (forthcoming). For simplicity, that model makes the extreme assumption that expectations are irrelevant in the absence of a promise. But the important point is that expectations of cooperation (or, in our case, reliance in anticipation of cooperation) matters more when a promise to cooperate was made to the recipient.

3 Design and Procedure

In this section, we describe the specifics of the design and procedures employed in our two experiments. In the first experiment, we asked subjects to report the likelihood they would punish, at a cost to themselves, non-cooperation by A in various iterations of this game. The scenarios they were reacting to were hypothetical and their responses had no effect on anyone's payoff including their own. In the second experiment, subjects observed the actual interactions of other subjects who had played a version of the trust game, and the decisions of all subjects, including subjects in the role of third-party observer, had monetary consequences.

3.1 Experiment 1: Vignette Study

In the first experiment, subjects were instructed to imagine that, like Player C, they had observed as third parties six iterations of the trust game in which B joined forces with A but A subsequently decided not to cooperate with B. There were three "Promise conditions" and three "No Promise conditions", each one characterized by a particular investment level chosen by B: 0, 3, and 6. In the Promise conditions subjects were asked to imagine that A had promised B that she would cooperate with him if he joined forces with her ("I promise that I will cooperate with you"), while in the No Promise conditions they were asked to imagine that A had told B that she planned on cooperating with him without making any promises ("All I can say is that I plan to cooperate with you, though I can't promise that I will do so"). Following the presentation of each scenario to subjects, they were asked to report the likelihood that they would choose to inflict a punishment on A for not cooperating with B at some small but not insignificant cost to themselves.[4]

Subjects saw six scenarios in a randomized sequence.[5] This allowed us to generate both

[4]We told subjects that inflicting punishment on A might consist of "boycotting A's business which results in a monetary loss for A." We also reminded them that "[i]nflicting this punishment, however, is not costless to you" as "boycotting a business that you have been buying from forces you to switch to another product requiring you to change your habits, pay more for the other product, consume a product that you like less, etc...)."

[5]The order was not completely random. Subjects either saw all three promise conditions first in a

between- and within-subject data. Subjects' responses to the first condition they saw constitute between-subject data. Subjects' responses in their entirety constitute within-subject data.

We programmed the vignettes using Qualtrics and recruited 1200 subjects from Amazon Mechanical Turk's pool of MTurk workers who had a HIT (Human Intelligence Task) approval rate of 95% or greater. We determined our sample size using a simulation based on pilot data on 300 subjects.[6]

Subjects were asked control questions to ascertain whether they had read and understood the scenario, and they were not allowed to proceed until they answered those questions correctly. At the end of the survey subjects were asked several other questions to assess how carefully and honestly they responded to the questions. We also elicited subjects' demographic characteristics.

Before subjects were presented with the scenarios, they were informed that they would be paid \$1.50 for participating and that the task would take approximately 10 to 15 minutes. The announced hourly wage was therefore \$6 to \$9 per hour. Thus, on average subjects could expect to receive more than the current federal minimum wage (\$7.25 per hour) and expected payments were much higher than the wages MTurk workers typically earn.[7] On average our subjects took 6.8 minutes to complete the task. Thus, the effective average hourly wage was \$13.24. We reproduce screenshots from the experiment in Appendix B.

3.2 Experiment 2: Study with Real Interactions and Incentives

In the second experiment, each subject was randomly and anonymously matched with two other subjects to play a version of the game described in Section 2. Two subjects from each group were assigned to be Player A and Player B and played a version of the trust game

randomized order and then all three no promise conditions in a randomized order or vice versa.

[6]We excluded the pilot data from our subsequent analysis. However, including the data does not change the statistical significance of any of our results.

[7]Studies have found a median hourly wage of \$1.38 (Horton & Chilton, 2010) and a typical payment of \$0.01-\$0.10 per HIT (Mason & Watts, 2010).

with four investment levels (zero to three).[8] The third subject in the group was assigned to be Player C, and so had the opportunity to punish A by reducing A's payoff at the end of the game at a monetary cost to himself. More specifically, by reducing his own payoff by one unit, C could reduce A's payoff by three units up to a maximum of 5 and 15 units respectively. Subjects were made aware that C would have the opportunity to do reduce A's payoff in this way at the end of the game.

We relabeled the actions when presenting the game to subjects to make the language as neutral as possible, in case the more descriptive language used in the first experiment was inadvertently suggestive. Thus, B had the option to simply "Exit" or "Enter" at the outset of the game. If B entered, he then had to choose an integer from a set, each one corresponding to the investment levels in the original trust game but not described as such. A's choices were described as simply "Left" (the cooperative action) or "Right" (the non-cooperative action). And C then had decide by how much he wanted to reduce A's payoff at a cost of a lesser reduction of his own payoff. Thus, his decision was not presented as involving punishment.[9]

The messages that A could choose between at the outset of the game were different from the messages that subjects were asked to imagine that A had made in the first experiment. A could either make a promise to cooperate with B by sending the message "I promise to choose Left" or remain silent. Thus, when A made no promise, she didn't communicate with B at all (as opposed to stating that she intended to cooperate with B while explicitly disclaiming any promise to that effect). If A sent the message, B received it immediately, and prior to making his punishment decision C was told that "A promised to choose Left." If A instead remained silent, B was simply told that "Player A has not sent you a message. "

[8] In giving them a choice among four investment levels, we kept our costs down as well as the number of questions that subjects needed to answer. Subjects assigned to be C answered 16 questions over three screens. If instead they had been asked about seven investment levels, they would have had to answer 28 questions.

[9] The prompt subjects assigned to be C saw was as follows: "You can now decide how much to spend on reducing A's payoff. Each token that you spend will reduce your payoff by 1 token and A's payoff by 3 tokens. If [randomly inserted option], how much would you like to spend reducing A's payoff?"

and C was told that "A did not make a promise" prior to making his punishment decision.[10]

We elicited the decisions of subjects assigned to be A or B using the direct response method. We used the strategy method to elicit the punishment decisions of subjects assigned to be C, which meant that they made their decisions without knowing whether or not A made a promise during the communication stage and A's and B's subsequent decisions during the trust game. That is, we elicited their punishment decisions for each possible scenario that could emerge from the trust game. The actual punishment that was inflicted on A (and so the subjects' final payoffs) was the punishment that C chose for the scenario that corresponded to the actual scenario that A and B played out. In this way, we generated within-subject data showing subjects' reactions to all relevant permutations of the game.

We also generated between-subject data by ensuring that the first scenario that subjects assigned to be C reacted to was randomly assigned across those subjects, and eliciting their reactions to that scenario before they learned that we would also ask them to indicate how they wanted to punish A in the other possible scenarios. Our aim here was to simulate a direct response to the scenario they were first randomly presented with.[11] Because of the number of possible combinations, we only randomized across four options: (1) promise, investment of 0, no cooperation; (2) promise, investment of 2, no cooperation; (3) no promise, investment of 0, no cooperation; and (4) no promise, investment of 4, no cooperation.[12]

After the main game was concluded, every subject who had been assigned to be A or B

[10]We designed the communication stage in this way because we thought that had subjects been allowed to engage in free-form communication, those in role A who didn't want to promise to cooperate with B probably would have done so by remaining silent or not mentioning promising at all rather than making a statement of intent that disclaimed a promise. We also thought that the conjunction of results from the two experiments would be more convincing if we could confirm our hypotheses using two different constructions of the message space.

[11]A concern about the strategy method is that subjects may experience it as artificial as they don't know what the others actually chose when responding and also therefore don't know which of their responses will end up being implemented. Brandts and Charness (2011) survey studies that employ both the direct response method and strategy method and, though they find that "there are significantly more studies that find no difference across elicitation methods than studies that find a difference" (p. 387), they also note that there is some evidence that emotions run higher when decisions are elicited under the direct-response method, perhaps because those decisions feel more real to subjects.

[12]We chose an investment level of 0 as a baseline and we chose an investment of 2 instead of 1 because 1 is the efficient investment level, a fact that might have influenced subjects' punishment decisions independently of the variables we are interested in.

was asked to report their perceptions of the appropriateness of various actions. We elicited their perceptions by employing the method developed by Krupka and Weber (2013), which is designed to elicit subjects' beliefs about relevant shared norms rather than their subjective attitudes. Thus, we asked subjects to answer on a four-point scale (very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate) while also telling them that they would receive \$0.10 every time they selected the *modal* response given by subjects who had been assigned the same role in the game.[13] Thus, subjects maximized their monetary payoff by selecting the action that the greatest number of similarly situated subjects also selected, rather than offering their own subjective views about the appropriateness of A's action. If a shared norm governing the appropriateness of A's actions exists, it should provide a focal point that (in the absence of any other focal point) subjects can use to solve the coordination game. This in turn suggests that subjects' responses will track the structure of any such norm.

The experiment was programmed in oTree and conducted in January 2020. We recruited 238 subjects from Amazon Mechanical Turk's pool of Master Workers who had a HIT approval rate of 98% or greater.[14]

Subjects were asked detailed control questions to determine whether they had understood the scenario, and they were not allowed to proceed until they had answered those questions correctly. Before subjects began the experiment, they were informed that the task would take approximately 15-20 minutes and that they would receive a \$1 fee for correctly answering the questions with the potential for a substantial performance bonus.[15] On average, subjects

[13]More specifically, for each scenario, each subject was asked to indicate whether she believes "that A's choice of Right would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate." She was then told that we would "determine which response was selected by the most participants assigned the role of Player A," and that if she gave "the response that is most frequently given by other people" she would "receive a bonus of \$0.10." The modal response was calculated for each group. So dictators were rewarded for giving the modal response of other dictators and recipients were rewarded for giving the modal response of other recipients.

[14]We ended up with data on 84 subjects in role C, 78 in role A, and 76 in role B. We don't have even numbers of each because subjects can drop out. As soon as the players were matched, those in role C could complete the game even if the subjects they were matched with dropped out. Subjects in roles A and B depended on the others and a drop out by one could affect the others.

[15]The exact language was: "You will receive USD 1 for participating. You have the opportunity to earn a substantial bonus." We did not announce an hourly wage.

took 11.7 minutes to complete the experiment so that the effective hourly wage was $21. The performance bonus depended on the total number of tokens they had at the end of the game. Each subject started with an endowment of tokens: 7 if assigned to be A; 7 if assigned to be B; and 15 if assigned to be C. Payoffs from the game were made in terms of tokens and so their total number of tokens increased or decreased depending on the outcome of the game. One token was worth $0.30, but if a subject's final token balance was negative, her bonus payment was zero. The game tree and a link to the full instructions was included at the top of every slide that subjects saw. We reproduce screenshots from the experiment (without repeating the game tree on each slide) in Appendix C.

4 Results

In this section, we describe in detail the results from both experiments. When testing our hypotheses with the within-subject data we use the Wilcoxon sign-rank test. When testing our hypotheses with the between-subject data we primarily use the Wilcoxon rank-sum test.[16]

4.1 Experiment 1

Table 1 and Figure 2 summarize the mean reported likelihood of punishment by treatment condition in experiment 1 for both sets of data.[17] Descriptively both our within-subject and

[16]The Wilcoxon rank-sum test is unavailable for testing H1.3 and H2.3. This is because testing for an interaction effect requires us to compare the difference in the reported willingness to punish or amount of punishment (or perceptions of appropriateness) for different reliance levels in the Promise and the No Promise conditions. The Wilcoxon ranksum test allows us to test for differences between unmatched data but not differences in differences.

[17]We excluded the responses of the 126 subjects who reported in the post-experiment survey that they did not or only "kind of" understood how A's and B's actions affected their payoffs. We also excluded the responses of four subjects who reported that they had done the study before. We have no good explanation as to why four subjects reported having taken the survey before. We provided links to participants which were only good for a single log in. We implemented filters preventing subjects (as identified by their MTurk IDs) from participating who had participated in pilots of our experiment or similar experiments we had run in the past. So the only explanation for the four self-reported repeat takers could be that subjects have multiple MTurk IDs or mistakenly checked the wrong box. Including these data does not, however, change the results. Similarly, results do not qualitatively change when we exclude subjects who report that they did not "take the scenario seriously," that they did not "carefully read the instructions," and that they chose their answers to make themselves "seem like a good person."

between-subject data are in large part line with our hypotheses. In line with hypothesis H1.1, subjects reported a higher average likelihood of punishing A when they were told that A made a promise to cooperate with B. Thus, the Promise lines in Figure 2 lie above the No Promise lines. In line with hypothesis H1.2, subjects also reported a higher average likelihood of punishing A the greater was B's investment level in both the Promise and No Promise conditions (with the exception of the (3-6) interval in the No Promise conditions in the between-subject data). Thus, both lines in Figure 2 are upward sloping (at least over the entire range). Finally, in line with hypothesis H1.3, the increase in the average reported likelihood of punishing A caused by higher investment is larger when subjects were told that A made a promise to cooperate with B. Thus, the Promise lines in Figure 2 are steeper than the No Promise lines over each interval and over the entire range.

Table 1: Mean Reported Willingess to Punish

	Observed Investments in the Relationship		
	0	3	6
Between-Subject			
No Promise	2.47 (N=173)	2.90 (N=183)	2.84 (N=183)
Promise	2.85 (N=156)	3.73 (N=194)	3.89 (N=186)
Within-Subject (N=1075)			
No Promise	1.92	2.55	2.76
Promise	2.75	3.81	4.11

Tests on the within-subject data yield support for all of our hypotheses.[18] Tests on the between-subject data find a promise per se effect (H1.2) for each level of investment.[19] Subjects report they are more willing to punish non-cooperation when it breaks a promise to cooperate. Tests on the between-subject data also suggest that reliance matters to subjects' reported willingness to punish if there was a promise,[20] and for the (0-3) and (0-6) investment

[18]We generally obtain significance at the 1% level. Only under a very restrictive definition of our data pool (eliminating all observations that might be invalid in the light of our post-experiment survey) do we find $p = 0.02$.

[19]Each is significant at the 1% level.

[20]The effect is significant at the 1% level for the (0-3) and (0-6) intervals, and at the 5% level ($p = 0.02$) for the (3-6) interval.

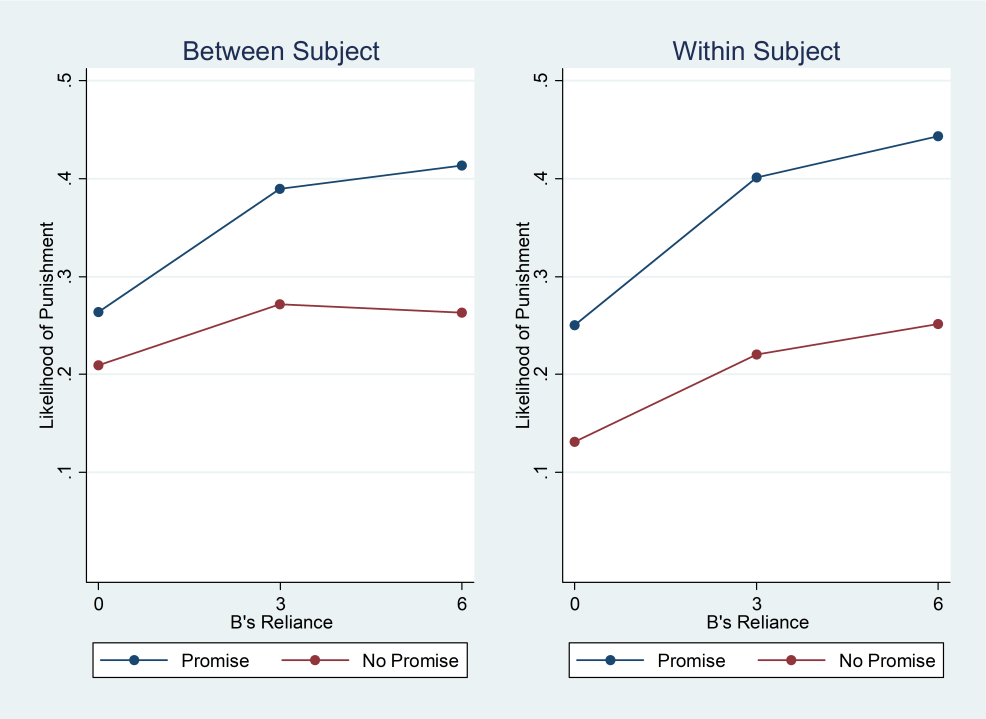


Figure 2: Average Reported Likelihood of Punishment Across Conditions

intervals if there was no promise.[21] Thus, the between-subject data support the reliance per se effect (H1.2) for the most part. Finally, we perform a bootstrapping procedure to see whether the between-subject data exhibit an interaction effect (H1.3). This yields significance at the 1% level for the (0-6) interval.[22] This suggests that increased reliance has a greater effect on the reported willingness of third parties to punish non-cooperation when non-cooperation breaks a promise to cooperate.

4.2 Experiment 2

Table 2 and Figure 3 summarize the average punishment by treatment by treatment condition in experiment 2 for both sets of punishment data. Approximately 53 percent of subjects assigned to role C inflict a positive amount of punishment on A.[23]

Descriptively the within-subject punishment data support our hypotheses for the most

[21]The effect is significant at the 1% level. For the (3-6) interval, mean effort goes down, contrary to H1.2, though the effect is not significant at any level (p=0.63).

[22]The bootstrapping procedure simulates synthetic samples and is described in detail in Appendix A.

[23]Approximately 80% of subjects assigned to role A make a promise, 89% of subjects assigned to role B enter the game, and of those 33% invest zero, 48% invest 1, 11% invest 2, and 8% invest 3.

part. The Promise line is above the No Promise line, suggesting that subjects punish non-cooperation more harshly when a promise was made regardless of the investment level (H1.1). Both lines are increasing as investment increases from zero to 1, and 2 to 3, and from 0 to 3 suggesting that higher investment increases their willingness to punish non-cooperation (H1.2). The Promise line is also increasing slightly between 1 and 2, though the No Promise line decreases slightly over that interval. Finally, the No Promise line is rising more steeply over each interval and the entire range when there was a Promise in support of the interaction effect (H1.3).

Descriptively the between-subject data support the existence of the interaction effect (H1.3). The Promise line rises steeply as investment rises from zero to 2, while the No Promise decreases. Thus, the data are also in line with the reliance per se effect (H1.2) when there was a promise, though not when there was no promise. But at investment to zero, subjects punish on average less harshly when there was a promise, contrary to the promise per se effect (H1.1).

Table 2: Mean Punishment Imposed

	Observed Investments in the Relationship			
	0	1	2	3
Between-Subject				
No Promise	0.52 (N=21)		0.41 (N=22)	
Promise	0.24 (N=21)		1 (N=20)	
Within-Subject (N=84)				
No Promise	0.77	1.01	1.04	1.23
Promise	0.33	0.52	0.51	0.61

A comparison of the within-subject and between-subject lines in Figure 3 shows that the reversal of the promise per se effect at zero investment in the between-subject data is driven mainly by subjects who are first presented with the scenario in which A promised B she would cooperate with him and B subsequently invested zero. Average punishment falls from nearly 0.8 in the within-subject data to just above 0.2 in the between-subject data. (Average punishment in the No Promise treatment at zero investment is also higher in the

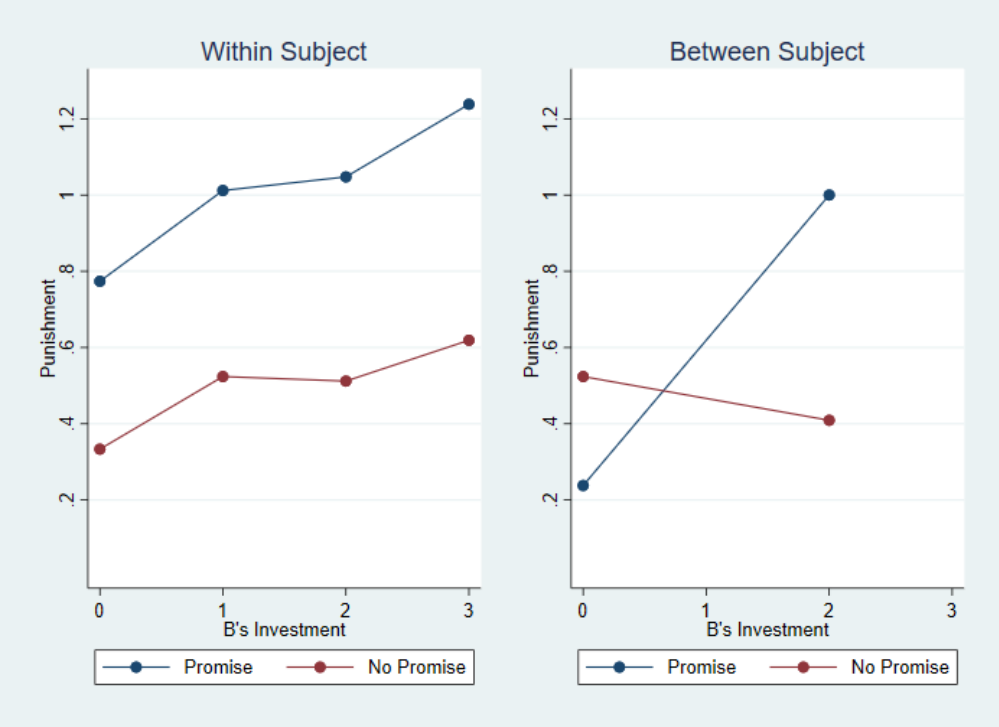


Figure 3: Average Reported Likelihood of Punishment Across Conditions

between-subject data but the difference is approximately 0.2 rather than 0.6.) This suggests that the absence of investment is particularly salient to subjects when they respond to this scenario first. Perhaps they think to themselves (consistent with the promise per se effect) that while breaking a promise is bad, it's really not so bad to break it when B communicated his distrust of A by investing zero.

A deeper dive into the data reveals evidence in line with this conjecture. Figure 4 shows that the within-subject responses of subjects who were exposed to the no-promise-zero-investment scenario first exhibit a promise per se effect, even though their between-subject responses cause the reversal in the between-subject data. That is, when exposed to the other conditions these subjects punish non-cooperation that breaks a promise more harshly than non-cooperation that doesn't even when investment is zero. For these subjects, moreover, there is a sharp uptick in the average punishment at investment level one, after which punishment remains relatively stable in both the Promise and No Promise conditions, suggesting that it is the presence or absence of investment that these subjects end up caring

about. In other words, it seems that the absence of investment is made particularly salient to subjects when they see the no-promise-zero-investment scenario first. Figure 4 shows that the within-subject responses of all four between-subject groups exhibit promise per se effects across the entire range of possible investment levels.

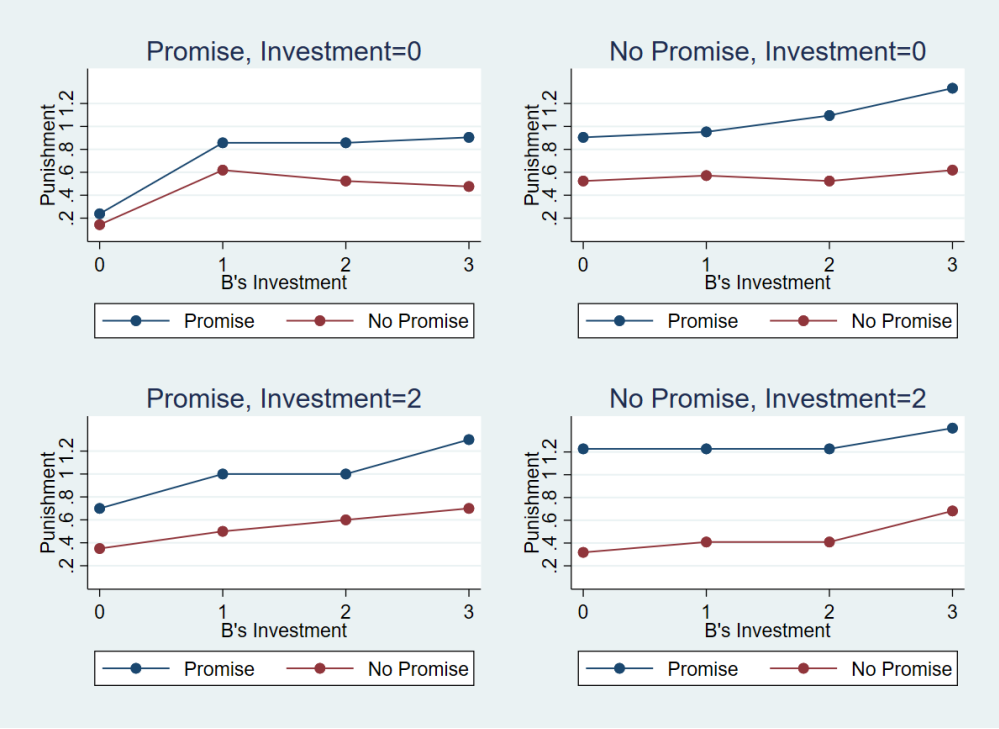


Figure 4: Within Subject Third-Party Punishment Decisions Broken Down By Between Subject Group (Incentivized)

Tests on the within-subject data confirm the promise per se effect (H1.1) for all four investment levels suggesting that subjects are willing to inflict a higher punishment on A when she made a promise regardless of the investment level.[24] Tests also confirm the reliance per se effect (H1.2) when a promise was made for all investment intervals except that between 1 and 2,[25] but only for investment increases between 0 and 1, 0 and 2, and 0 and 3 when there was no promise.[26] Finally, the interaction effect (H1.3) is confirmed for an increase in investment from 1 to 3,[27] but not for other increases.

[24] In each case, the effect is significant at the 1% level.
[25] In each case, the effect is significant at the 1% level.
[26] $p = 0.01$, $p = 0.03$, and $p < 0.01$ respectively.
[27] $p = 0.03$.

Tests on the between-subject data don't confirm the promise per se effect (H1.1) for either investment level. We find support for the reliance per se effect (H1.2) when there is a promise though only at the 10% level,[28] but not where there is no promise (indeed, the effect goes in the wrong direction). But these two results together offer indirect support for the interaction effect (H1.3). (Because between-subject data are unmatched, a bootstrapping procedure that generates artificial samples, which we haven't yet performed, is required to directly test the interaction effect.)

Tables 3 and 4 and Figures 5 and 6 illustrate perceptions of social appropriateness reported by subjects in role A and subjects in role B respectively. Descriptively, they exhibit similar patterns to the punishment data consistent with our Hypothesis 2 that propensities to punish are driven by the same underlying promissory norms.

Table 3: Perceptions of Inappropriateness (Player A)

	Observed Investments in the Relationship			
	0	1	2	3
Between-Subject				
No Promise	2.40 (N=20)		1.95 (N=19)	
Promise	2.50 (N=20)		3.00 (N=19)	
Within-Subject (N=78)				
No Promise	1.83	1.92	2.06	2.17
Promise	2.47	2.53	2.62	2.56

Table 4: Perceptions of Inappropriateness (Player B)

	Observed Investments in the Relationship			
	0	1	2	3
Between-Subject				
No Promise	2.81 (N=16)		2.05 (N=20)	
Promise	2.77 (N=22)		3.00 (N=18)	
Within-Subject (N=76)				
No Promise	1.97	1.98	2.17	2.18
Promise	2.39	2.21	2.34	2.53

[28] $p = 0.73$.

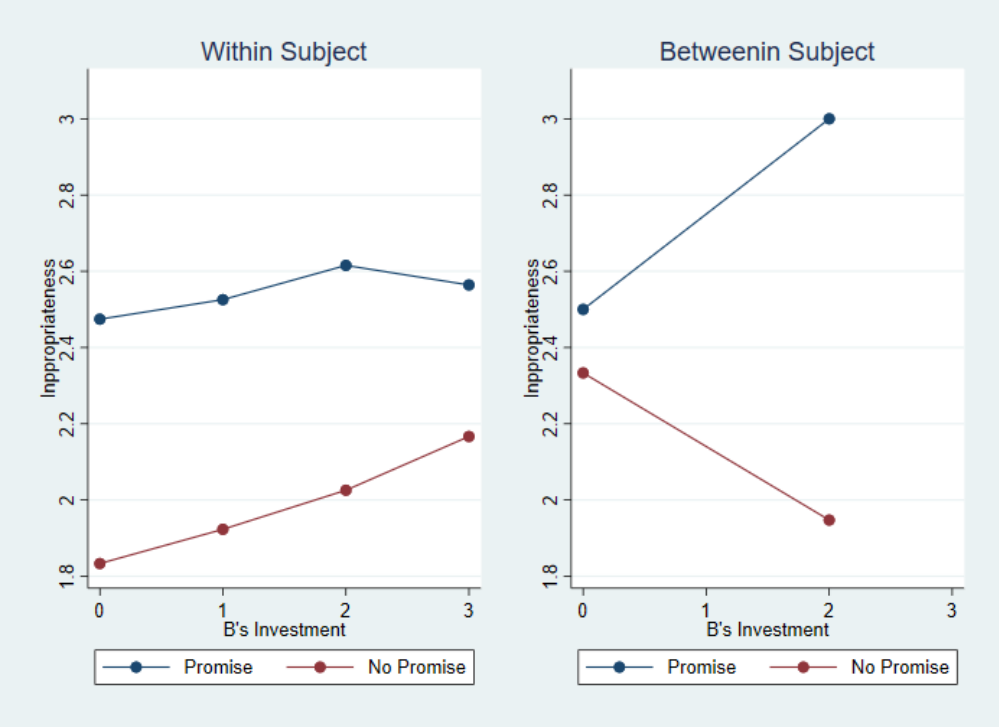


Figure 5: Player A's Perceptions of Inappropriateness Across Conditions (Incentivized)

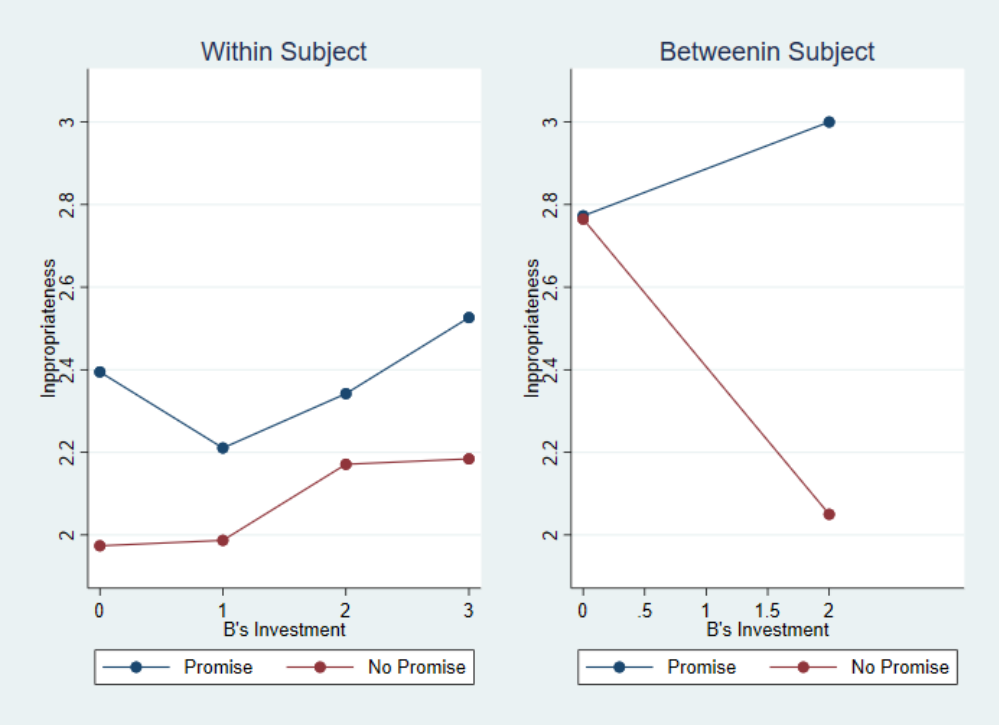


Figure 6: Player B's Perceptions of Inappropriateness Across Conditions (Incentivized)

The promise per se effect (H2.1) is significant for the within-subject appropriateness

data elicited from subjects in role A for all investment levels,[29] and for investment level 2 for the between-subject data.[30] It is significant for the the within-subject data elicited from subjects in role B for investment levels 0 and 3,[31] and for investment level 2 for the between-subject data.[32] As with the punishment data, we see a disappearance of the promise per se effect in the between-subject data for subjects in both roles when investment is zero, but in this case, the disappearance seems to be driven more by increased perceptions of social inappropriateness when A made no promise than by decreased perceptions of social inappropriateness when A made a promise and B invested zero.

The effect of increased investment in the absence of a promise on A's perceptions of appropriateness in the within-subject data is significant at 1% for the increase from 0 to 3. for the increase from 0 to 1, 0 to 2, and 0 to 3 at 10%, 5%, and 1% respectively.[33] When a promise was made, however, none of the differences are significant in the within-subject data. In the between-subject data, by contrast, the effect of higher investment on A's perceptions of appropriateness when a promise was made is significant at 10%.[34] Thus, we have support for the reliance per se effect (H2.2) in the absence of a promise for the within-subject data and when a promise was made for the between-subject data, but not for the within-subject data when a promise was made, nor for the between-subject data in the absence of a promise. In the between-subject data, there is indirect support for an interaction effect (H2.3), given the significant increase in inappropriateness perceptions only when A made a promise.

B's perceptions of appropriateness exhibit a similar pattern in response to increased reliance. In the within-subject data, the effect of an increase from 0 to 3 is significant at 10% when A made a promise.[35] When A made no promise, the effect of an increase in

[29] $p < 0.01$ for investment levels 0 to 2 and $p = 0.02$ for investment level 3.

[30] $p < 0.01$.

[31] $p < 0.01$ in both cases.

[32] $p = 0.01$.

[33] The increase from 0 to 1 is significant at 10% ($p = 0.05$), and the increase from 0 to 2 at 5% ($p = 0.04$). The increase from 1 to 3 is significant at 5% ($p = 0.02$) and the increase from 2 to 3 is significant at 10% ($p = 0.07$).

[34] $p = 0.07$.

[35] $p = 0.08$. The effect of increasing investment from 1 to 3 is significant at 1% and from 2 to 3 at 10% ($p = 0.06$).

investment from 0 to 3 is significant at 5%.[36] Thus, the within-subject data offer some support for H2.2, though no convincing evidence of the interaction effect.[37] In the between-subject data, the increase when a promise was made is not significant, but the *decrease* when a promise was not made is significant at 10%,[38]. So the between-subject data don't support the reliance per se effect (H2.2), but they do offer some indirect support for the interaction effect (H2.3).

5 Discussion

Stone and Stremitzer (forthcoming) found evidence that the increased willingness of a promisor to keep a promise on which the promisee had relied more created an incentive for promisees to strategically overinvest in reliance in order to psychologically lock the promisor in to keeping her promise. This overinvestment disappeared when the promise was enforced with expectation damages, as the promisee could then rely on the legal regime rather than overinvestment to motivate the promisor to keep her promise. This suggests that legal enforcement not only has the potential to mitigate problems of underinvestment—the focus of the literature on breach remedies and the hold-up problem.[39] It may also reduce overinvestment arising from psychological lock-in.

Here we find evidence that third-party observers are more willing to punish a non-cooperative player A more when A promised B that she would cooperate. Our evidence also suggests that this willingness to punish increases with the investment B made in the relationship creating another channel for lock-in. In contrast to the mechanism posited in Stone and Stremitzer (forthcoming), such a reputation-based lock-in mechanism doesn't rely on the internal moral motivations of the promisor. Thus, promisees should be able to har-

[36] $p = 0.04$. For an increase from 0 to 2, $p = 0.05$, and for an increase from 1 to 3, $p < 0.01$

[37] Except for the increase from 2 to 3 where the interaction effect is significant at 10% ($p = 0.05$).

[38] $p = 0.05$.

[39] On breach remedies in particular, see, e.g., Shavell (1980, 1984), Rogerson (1984), Cooter and Eisenberg (1985), Edlin and Reichelstein (1996), Edlin (1996), Che and Chung (1999), Schweizer (2006) , Ohlendorf (2009), Stremitzer (2012). On the hold-up problem generally, see, e.g., Williamson (1979, 1985), Grout (1984), Grossman and Hart (1986), Hart and Moore (1988), Chung (1991), Aghion, Dewatripont, and Rey (1994), Noeldeke and Schmidt (1995), Che and Hausch (1999).

ness this reputation-based lock-in mechanism by increasing their reliance investments, and so may have an incentive to overinvest even if faced with a self-interested promisor. Notice also that we find that there is something distinctively promissory about this effect, insofar as our results suggest that the willingness to punish increases more with B's investment level if A made a promise.

From a methodological standpoint, the fact that the results of our incentivized experiment (experiment 2) are roughly in line with those of our vignette study (experiment 1) suggests that subjects reported willingness to punish non-cooperative behavior reliably tracks in important structural respects their actual willingness to punish when faced with pecuniary incentives to do otherwise. This suggests that the vignette method may be a valid method of uncovering the structure of underlying norms, even if it doesn't provide such compelling evidence of the overall willingness of persons to actually engage in altruistic punishment in accordance with those norms when faced with an incentive to do otherwise.

6 Conclusion

We have provided experimental evidence that third parties observing an interaction between a promisor and promisee are more willing to punish non-cooperative behavior when it breaks a promise to cooperate than in the absence of any such promise. They are also more willing to punish non-cooperative behavior when the recipient relied on an expectation of cooperation. And this effect of higher investment on the willingness of third parties to punish is greater when there was a promise of cooepration. In these ways, third parties seem to be sensitive to the same underlying promissory norms that motivate promisors to keep their promises: they view breaking a promise as morally worse than a mere failure to cooperate, and so as more deserving of punishment; and they view promise breaking as morally worse the more the promisee has relied on it. These results suggest that reputational forces may provide a promisee with a mechanism for locking a promisor who isn't herself motivated by promissory norms into keeping her promise: by investing more, the promisee increases the chance that third parties who are sensitive to promissory norms will punish the promisor should she

break the promise.

Our results also highlight a more general point. Studying the reasons why people voluntarily keep their promises sheds light on the dynamics of altruistic punishment, and, we conjecture, the dynamics second-party enforcement mechanisms that are embedded in relational contracts. The promissory norms that motivate many promisors to keep their promises also inform the behavior of those who can and are able to punish promise breakers. This suggests that there is a stronger link between relational contracting, altruistic punishment, and voluntary promise keeping than has been noticed in the contract theory literature to date.

References

Aghion, P., M. Dewatripont, and P. Rey. 1994. "Renegotiation Design with Unverifiable Information." *Econometrica*, 62: 257–282.

Brandts, J.;, and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons." *Experimental Economics*, 14(3)): 375–398.

Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership." *Econometrica*, 74: 1579–1601.

Charness, Gary, and Martin Dufwenberg. 2010. "Bare Promises: An Experiment." *Economics Letters*, 107: 281–83.

Che, Y. K., and D. B. Hausch. 1999. "Cooperative Investments and the Value of Contracting." *American Economic Review*, 89: 125–147.

Che, Y. K., and T. Y. Chung. 1999. "Contract Damages and Cooperative Investments." *RAND Journal of Economics*, 30: 84–105.

Chung, T. Y. 1991. "Incomplete Contracts, Specific Investments, and Risk Sharing." *Review of Economic Studies*, 58: 1031–1042.

Dufwenberg, Martin, and Uri Gneezy. 2000. "Measuring Beliefs in and Experimental Lost Wallet Game." *Games and Economics Behavior*, 30: 163–182.

Ederer, Florian, and Alexander Stremitzer. 2017. "Promises and Expectations." *Games and Economic Behavior*, 106: 161–178.

Edlin, Aaron S. 1996. "Cadillac Contracts and Up-front Payments: Efficient Investment Under Expectation Damages." *Journal of Law, Economics, and Organization*, 12: 98–119.

Edlin, Aaron S., and Stefan Reichelstein. 1996. "Holdups, Standard Breach Remedies, and Optimal Investment." *American Economic Review*, 86: 478–501.

Ellingsen, Tore, and Magnus Johannesson. 2004. "Promises, Threats and Fairness." *Economic Journal*, 114: 397–420.

Fehr, Ernst, and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior*, 25: 63–87.

Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr. 2005. "1 Moral Sentiments and Material Interests: Origins, Evidence, and Consequences." *Moral sentiments and material interests*, 1.

Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership - A Theory of Vertical and Lateral Integration." *Journal of Political Economy*, 94: 691–719.

Grout, Paul A. 1984. "Investment and Wages in the Absence of Binding Contracts: A Nash Bargaining Approach." *Econometrica*, 52: 449–460.

Hart, Oliver, and John Moore. 1988. "Incomplete Contracts and Renegotiation." *Econometrica*, 56: 755–785.

Hart, Oliver, and John Moore. 2008. "Contracts as Reference Points." *The Quarterly Journal of Eeconomics*, 123: 1–48.

Horton, John J., and Lydia B. Chilton. 2010. "The Labor Economics of Paid Crowdsourcing." *EC '10*, 209–218. New York, NY, USA: ACM.

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11: 495–524.

Mason, Winter, and Duncan J. Watts. 2010. "Financial Incentives and the Performance of Crowds." *ACM SigKDD Explorations Newsletter*, 11: 100–108.

Mischkowski, Dorothee, Rebecca Stone, and Alexander Stremitzer. forthcoming. "Promises, Expectations, and Social Cooperation." *Journal of Law and Economics*.

Nöldeke, G., and K. M. Schmidt. 1995. "Option Contracts and Renegotiation - a Solution to the Hold-Up Problem." *RAND Journal of Economics*, 26: 163–179.

Ohlendorf, Susanne. 2009. "Expectation Damages, Divisible Contracts, and Bilateral Investment." *American Economic Review*, 99 (4): 1608–1618.

Rogerson, William P. 1984. "Efficient Reliance and Damage Measures for Breach of Contract." *RAND Journal of Economics*, 15: 37–53.

Schweizer, U. 2006. "Cooperative Investments Induced by Contract Law." *RAND Journal of Economics*, Spring 37-1: 134–145.

Selten, R. 1967. "Die Strategiemethode zur Erforschung des eingeschraenkt rationalen Verhaltens im Rahmen eines Oligopolexperiments." 136–168. in *Beitraege zur experimentellen Wirtschaftsforschung*, edited by H. Sauermann., Tuebingen: Mohr.

Shavell, S. 1980. "Damage Measures for Breach of Contract." *Bell Journal of Economics*, 11: 466–490.

Stone, Rebecca, and Alexander Stremitzer. forthcoming. "Promises, Reliance, and Psychological Lock-in." *Journal of Legal Studies*.

Stremitzer, A. 2012. "Standard Breach Remedies, Quality Thresholds, and Cooperative Investments." *Journal of Law, Economics, and Organization*, 28(2): 337–359.

Vanberg, Christoph. 2008. "Why Do People Keep Their Promises? An Experimental Test of Two Explanations." *Econometrica*, 76: 467–1480.

Williamson, Oliver E. 1979. "Transaction-Cost Economics: The Governance of Contractual Relations." *Journal of Law and Economics*, 22: 233–61.

Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism.* New York: Free Press.

APPENDIX A: BOOTSTRAP

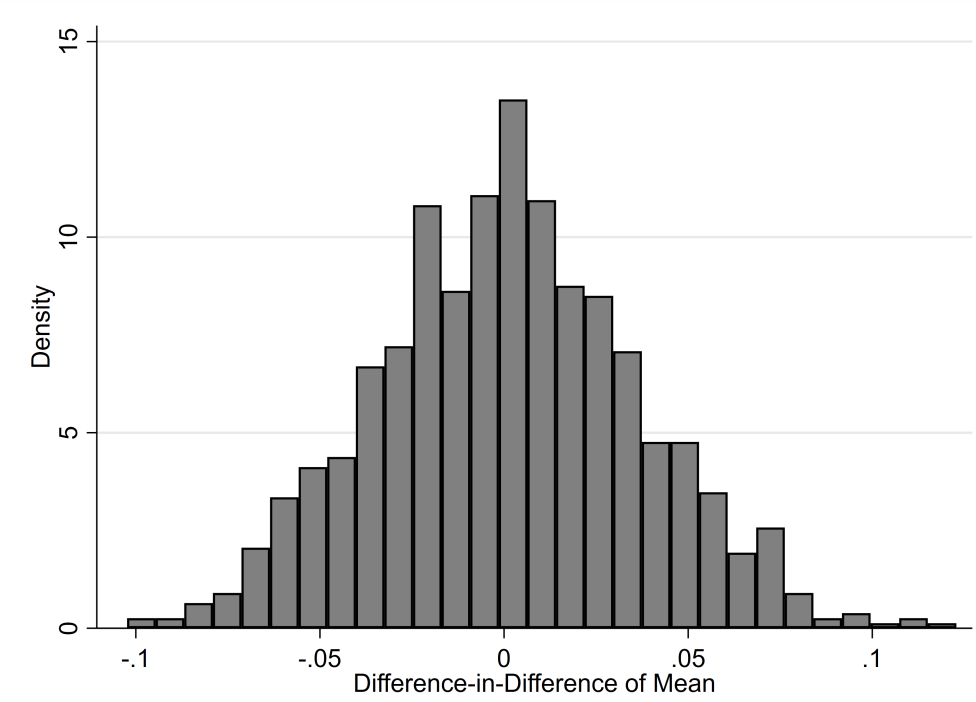


Figure 7: Simulated Distributions for the (0-6) Gap

We performed a two-step bootstrapping procedure in order to generate a non-parametric test of H1.4 on the (0-6) gap using the between-subject data from the first experiment. (We will eventually repeat this exercise using the between-subject data from the second experiment.) Let $\hat{\theta}_r$ and θ_r be the estimators of the mean reported likelihood of punishment in the Promise and No Promise conditions respectively when the promisee's reliance investments are r. We observe that $\left(\hat{\theta}_6 - \theta_6\right) - \left(\hat{\theta}_0 - \theta_0\right) = 0.097$. That is, the difference in means between the Promise and the No Promise samples is higher if promisees' reliance investments are 6 than if they are 0. We want to know the probability with which we would observe this positive difference-in-difference of means by chance. In other words, we want to test the null hypothesis that $\left(\hat{\theta}'_6 - \theta'_6\right) - \left(\hat{\theta}'_0 - \theta'_0\right) = 0$, where $\hat{\theta}'_1, \theta'_1, \hat{\theta}'_0, \theta'_0$ are the means of the underlying distributions from which our samples are drawn.

We tested the null hypothesis in two steps (see, e.g., Efron and Tibshrani, 1993, pp. 220-223). First, we recentered the original samples to conform with the null hypothesis. Specifically, we subtracted from each observation in each of the four samples the respective sample means and then added to each observation in the two Promise samples the mean effect of promising. In other words, if the mean for the combined No Promise samples is $\bar{x}$ and the mean for the combined Promise samples is $\bar{y}$, we added $(\bar{y} - \bar{x})$ to each observation in the two Promise samples.[40]

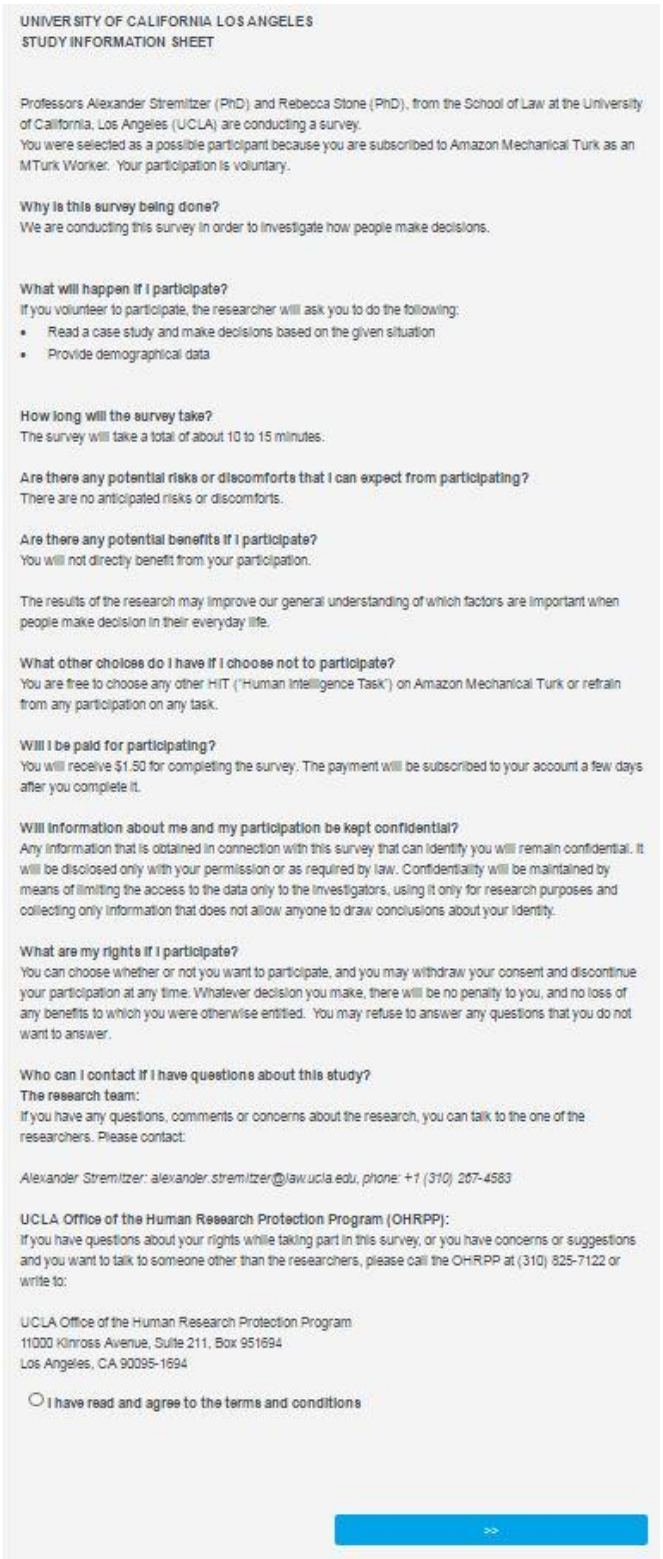
[40]By subtracting the sample means, we made our data conform to the hypothesis $\hat{\theta}'_6 = \theta'_6 = \hat{\theta}'_0 = \theta'_0 = 0$. In doing so, we eliminated all of our hypothesized effects from our data. By adding back $(\bar{y} - \bar{x})$ to the

We then created four synthetic samples with sample sizes equal to our real samples by randomly drawing with replacement from each of the four samples that were constructed above. We then calculated the difference-in-difference $\left(\hat{\theta}''_6 - \theta''_6\right) - \left(\hat{\theta}''_0 - \theta''_0\right)$, where $\hat{\theta}''_1, \theta''_1, \hat{\theta}''_0, \theta''_0$ are the means of these synthetic samples. After repeating this procedure 10,000 times, we obtained a simulated distribution of the differences-in-differences of the means that would arise if the null hypothesis were true (that is, if the difference of means between the Promise and the No Promise conditions was equal across different reliance levels).

The area under the curve to the right of the observed estimator 0.097 corresponds to the probability that a greater or equal difference-in-difference would have been observed if the null hypothesis were true. This value, 0.004, is small enough for us to reject the null hypothesis at the 1% level.

observations in the promise samples, we effectively added back in the expectations-independent effect of promising, so that our data ended up conforming to our less restrictive null hypothesis $\left(\hat{\theta}'_6 - \theta'_6\right) - \left(\hat{\theta}'_0 - \theta'_0\right) = 0$. We didn't add back in the effect of reliance alone, because doing so would leave this hypothesis unchanged.

APPENDIX B: INSTRUCTIONS (VIGNETTE)



UNIVERSITY OF CALIFORNIA LOS ANGELES
STUDY INFORMATION SHEET

Professors Alexander Stremitzer (PhD) and Rebecca Stone (PhD), from the School of Law at the University of California, Los Angeles (UCLA) are conducting a survey.
You were selected as a possible participant because you are subscribed to Amazon Mechanical Turk as an MTurk Worker. Your participation is voluntary.

Why is this survey being done?
We are conducting this survey in order to investigate how people make decisions.

What will happen if I participate?
If you volunteer to participate, the researcher will ask you to do the following:

- Read a case study and make decisions based on the given situation
- Provide demographical data

How long will the survey take?
The survey will take a total of about 10 to 15 minutes.

Are there any potential risks or discomforts that I can expect from participating?
There are no anticipated risks or discomforts.

Are there any potential benefits if I participate?
You will not directly benefit from your participation.

The results of the research may improve our general understanding of which factors are important when people make decision in their everyday life.

What other choices do I have if I choose not to participate?
You are free to choose any other HIT ("Human Intelligence Task") on Amazon Mechanical Turk or refrain from any participation on any task.

Will I be paid for participating?
You will receive $1.50 for completing the survey. The payment will be subscribed to your account a few days after you complete it.

Will information about me and my participation be kept confidential?
Any information that is obtained in connection with this survey that can identify you will remain confidential. It will be disclosed only with your permission or as required by law. Confidentiality will be maintained by means of limiting the access to the data only to the investigators, using it only for research purposes and collecting only information that does not allow anyone to draw conclusions about your identity.

What are my rights if I participate?
You can choose whether or not you want to participate, and you may withdraw your consent and discontinue your participation at any time. Whatever decision you make, there will be no penalty to you, and no loss of any benefits to which you were otherwise entitled. You may refuse to answer any questions that you do not want to answer.

Who can I contact if I have questions about this study?
The research team:
If you have any questions, comments or concerns about the research, you can talk to the one of the researchers. Please contact:

Alexander Stremitzer: alexander.stremitzer@law.ucla.edu, phone: +1 (310) 267-4583

UCLA Office of the Human Research Protection Program (OHRPP):
If you have questions about your rights while taking part in this survey, or you have concerns or suggestions and you want to talk to someone other than the researchers, please call the OHRPP at (310) 825-7122 or write to:

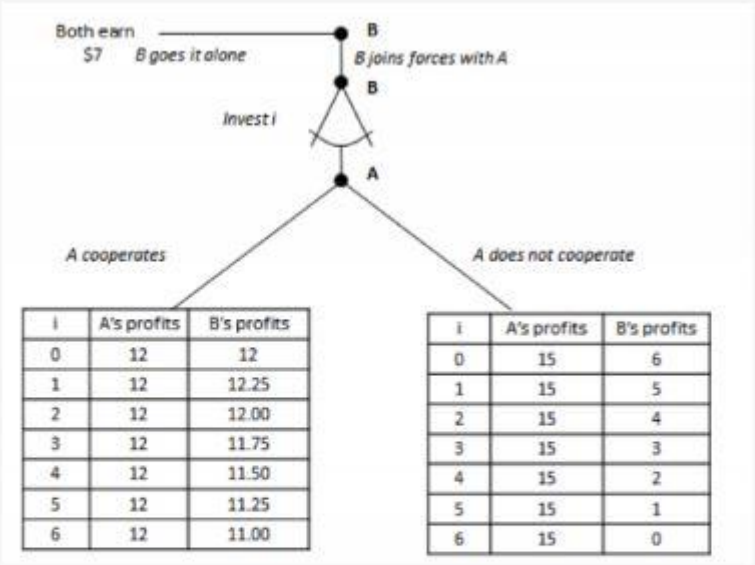
UCLA Office of the Human Research Protection Program
11000 Kinross Avenue, Suite 211, Box 951694
Los Angeles, CA 90095-1694

○ **I have read and agree to the terms and conditions**

>>

Thank you for participating in this survey. The purpose of this survey is to study how people make decisions in certain situations. You will earn $1.50 for completing the survey. The survey will take approximately 10-15 minutes.

>>



i	A's profits	B's profits
0	12	12
1	12	12.25
2	12	12.00
3	12	11.75
4	12	11.50
5	12	11.25
6	12	11.00

i	A's profits	B's profits
0	15	6
1	15	5
2	15	4
3	15	3
4	15	2
5	15	1
6	15	0

Imagine you witness the following:

Two people, A and B, have an opportunity to cooperate with each other.

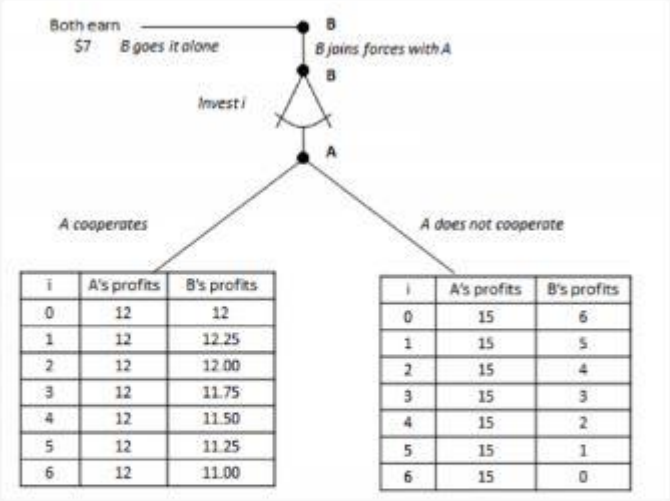
B must first choose between two alternatives (see diagram above): going it alone, or joining forces with A.

The upside of joining forces with A is that the combined monetary payoff of the two parties is much higher than if B were to go it alone.

The downside is that by joining forces with A, B puts himself at the mercy of A. This is because once B has decided to join forces with A, A can unilaterally decide not to cooperate with B in which case A gets all the profit from the venture. In other words, if B thinks that A is not going to cooperate, B is better off going it alone. However, that would leave both A and B worse off than if B had joined forces with A and A had cooperated.

If B decides to join forces with A, then, before A makes the decision whether to cooperate or not, B may invest in their cooperative venture. B's investment is lost if A does not cooperate. The higher B's level of investment, the higher will be B's loss (see right payoff table). However, if A cooperates with B, B's profits initially increase with his investment but then decrease once his investment rises beyond a certain point (see left payoff table).

>>



i	A's profits	B's profits
0	12	12
1	12	12.25
2	12	12.00
3	12	11.75
4	12	11.50
5	12	11.25
6	12	11.00

i	A's profits	B's profits
0	15	6
1	15	5
2	15	4
3	15	3
4	15	2
5	15	1
6	15	0

Preliminary Questions

1) What are the parties' profits when A cooperates and B has invested 3?

Profit of A:

Profit of B:

2) What are the parties' profits when A doesn't cooperate and B has invested 3?

Profit of A:

Profit of B:

3) If B thinks that A won't cooperate, which action will give B a higher profit?

- ○ Going it alone
- ○ Joining forces with A

>>

Imagine, you witness the following conversation:

B says to A:

"I would really like to join forces with you, but how can I be sure that you will cooperate with me rather than going it alone and taking all the profit for yourself?"

A responds:

"All I can say is that I plan to cooperate with you, though I can't promise that I will do so."

B listens carefully and decides to join forces with A.

A ***decides not to cooperate*** with B. Remember: A ***made no promise to cooperate*** with B. You observe this.

Imagine you can inflict a monetary punishment on A for not cooperating (e.g., by boycotting A's business which results in a monetary loss for A). Inflicting this punishment, however, is not costless to you (boycotting a business that you have been buying from forces you to switch to another product requiring you to change your habits, pay more for the other product, consume a product you like less, etc…).

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 3 knowing that A did not make a promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 0 knowing that A did not make a promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 6 knowing that A did not make a promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

Imagine, you witness the following conversation:

B says to A:

"I would really like to join forces with you, but how can I be sure that you will cooperate with me rather than going it alone and taking all the profit for yourself?"

A responds:

"I understand that you are worried I could take advantage of you, but I promise I will cooperate with you."

B says:

"Okay, if you promise to cooperate with me, let's work together."

B subsequently decides to join forces with A.

A ***decides not to cooperate*** with B. Remember: A ***promised to cooperate*** with B. You observe this.

Imagine you have the power to inflict a punishment on A for breaking the promise to B (e.g., by boycotting A's business which results in a monetary loss for A). Inflicting this punishment, however, is not costless to you (boycotting a business that you have been buying from forces you to switch to another product requiring you to change your habits, pay more for the other product, consume a product you like less, etc…).

>>

Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 6 in reliance on A's promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
○	○	○	○	○

>>

○ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 3 in reliance on A's promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
○	○	○	○	○

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 0 in reliance on A's promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

Did you understand how A's and B's actions affect their profits?

- ◯ Yes
- ◯ No
- ◯ Kind of

Finally, please answer the following questions:

	Yes	No
I didn't take the scenario seriously. I just wanted to earn the $1.50 fee as quickly as possible.	◯	◯
I carefully read the instructions.	◯	◯
I chose my answers in order to make myself seem like a good person.	◯	◯
This is the first time I have completed this survey.	◯	◯

>>

Please provide some demographical information.

What is your age?

What is your gender?

- ○ Female
- ○ Male

Is English your first language?

- ○ Yes
- ○ No

What is your highest level of schooling?

- ○ Master's, doctoral, or professional degree such as medicine or law
- ○ Bachelor's degree
- ○ Associate's degree
- ○ Vocational or technical certificate / diploma after high school (such as cosmetics)
- ○ High school diploma
- ○ I did not complete high school

Are you an mTurk Master Worker (your response to this questions will have no effect on your payout)?

- ○ Yes
- ○ No
- ○ I do not know what an mTurk Master Worker is.

>>

Thank you for participating in the survey.

Here is your MTurk Code: 752796

To receive payment for participating, click "Accept HIT" in the Mechanical Turk window, enter this code, and then click "Submit".

APPENDIX C: INSTRUCTIONS (INCENTIVIZED)

Instructions Page:

Instructions

Welcome to this experiment.

You will earn **$1** for carefully reading the instructions and answering questions to check your understanding. You can earn additional money based on your decisions and the decisions of the other participants.

Two people, A and B, have an opportunity to cooperate with each other. Before the game starts, A has the opportunity to send a message to B.

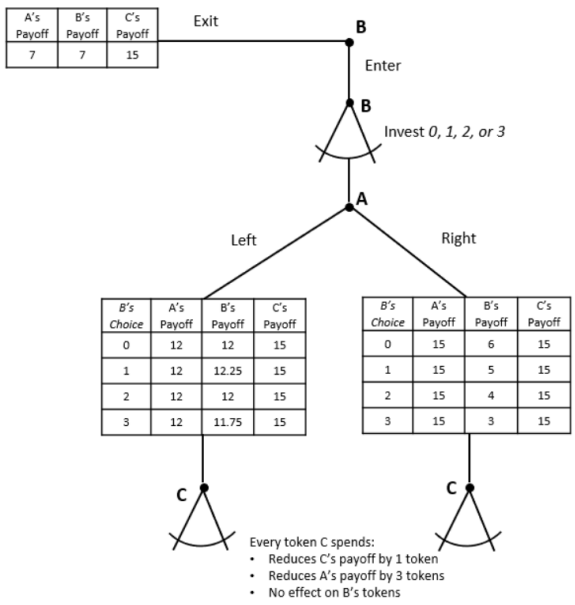
B must first choose between two alternatives (see diagram): Enter or Exit. The upside of entering is that B can earn more than by exiting (if A chooses Left). The downside is that B can also earn less (if A chooses Right).

If B decides to enter, B then makes a choice of 0, 1, 2, or 3. This choice only affects B, and the value of this choice to B depends on A's subsequent decision.

After B's decisions, A will choose between Left or Right, which will affect both A's and B's payoffs.

Finally C, who has observed the interaction between A and B, has the opportunity to reduce A's payoff. Doing so is costly to C but is even more costly for A.

The structure of the experiment is shown in the diagram. More detailed instructions are below.



Next

Time left to complete this page: **19:48**

Detailed Instructions

This experiment involves three participants: Player A, Player B, and Player C. After reading the instructions and answering some questions to check your understanding you will be randomly assigned to one of the three roles and matched with two other players. Your identity and the identity of the other two participants are anonymous. You will receive **$1** for reading the instructions and answering the preliminary questions. You can earn an additional bonus payment depending on your decisions and those of the other two players.

The payoffs for the experiment are in tokens. At the end of the experiment, the total amount of tokens earned will be converted and paid out to dollars at the rate of:

1 token = $0.30.

Please note that your tokens may be negative, in which case you will receive no bonus payment.

At the beginning of the game, players receive the following endowments:

- Player A: 7 tokens
- Player B: 7 tokens.
- Player C: 15 tokens.

Step 1: Communication Phase. Player A can choose whether or not to send a message to Player B.

Step 2: Player B's Decision to "Exit" or "Enter":

- If Player B chooses "Exit", **the game ends**. A and B each receive 7 tokens, and C receives 15 tokens.
- If Player B chooses "Enter", the game continues.

Step 3: Player B's Decision. Player B can invest **0, 1, 2, or 3.**

Step 4: Player A's Decision. Player A must decide between "Left" and "Right":

- If Player A chooses "Left", the preliminary payoffs are:

B's Choice	A's Payoff	B's Payoff	C's Payoff
0	12	12	15
1	12.25	11	15
2	12	12	15
3	12	11.75	15

- If Player A chooses "Right", the preliminary payoffs are:

Investment	A's Payoff	B's Payoff	C's Payoff
0	15	6	15
1	15	5	15
2	15	4	15
3	15	3	15

Step 5: Player C's Decision. Player C can spend tokens to reduce Player A's payoff. Each token spent reduces Player A's payoff by 3 tokens, while leaving B's payoff unchanged. Player C can choose to spend between 0 and 5 tokens.
The following table shows an example of how C's and A's payoffs are reduced depending on the number of tokens C spends.

Number of Tokens Spent by C	Reduction in C's Payoff	Reduction in A's Payoff
0	0	0
1	1	3
2	2	6
3	3	9
4	4	12
5	5	15

The final payoffs are given in the following tables:

- If Player B Enters, Player A chooses **Left** and Player C chooses to spend **x on reducing** A's payoff, payoffs are:

B's Choice	A's Payoff	B's Payoff	C's Payoff
0	12-3x	12	15-x
1	12-3x	12.25	15-x
2	12-3x	12	15-x
3	12-3x	11.75	15-x

- If Player B Enters, Player A chooses **Right**, and Player C chooses to spend **x on reducing** A's payoff, payoffs are:

B's Choice	A's Payoff	B's Payoff	C's Payoff
0	15-3x	6	15-x
1	15-3x	5	15-x
2	15-3x	4	15-x
3	15-3x	3	15-x

Next

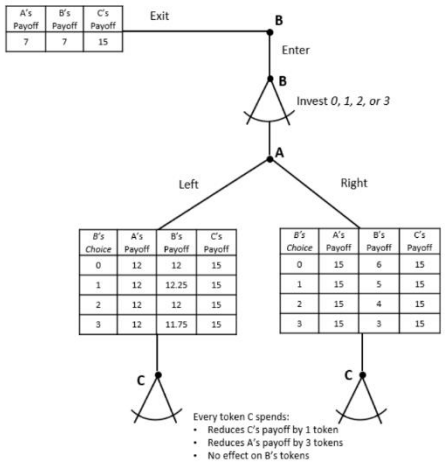
Preliminary Questions:

Before the experiment begins, you will be presented with some preliminary questions to check your understanding of how the decisions affect the players' payoffs. You must get the answers correct to proceed.

Next

Time left to complete this page: **7:40**

Preliminary Questions



1. What are the payoffs of Player A, Player B, and Player C if Player B enters; Player B chooses 2; Player A chooses Left; and Player C spends 0 on reducing A's payoff?

Player A's Payoff:

12

Player B's Payoff:

12

Player C's Payoff:

15

2. What are the payoffs of Player A, Player B, and Player C if Player B enters; Player B chooses 2; Player A chooses Left; and Player C spends 2 on reducing A's payoff?

Player A's Payoff:

6

Player B's Payoff:

12

Player C's Payoff:

13

3. What are the payoffs of Player A, Player B, and Player C if Player B enters; Player B chooses 2; Player A chooses Right; and Player C spends 2 on reducing A's payoff?

Player A's Payoff:

9

Player B's Payoff:

4

Player C's Payoff:

13

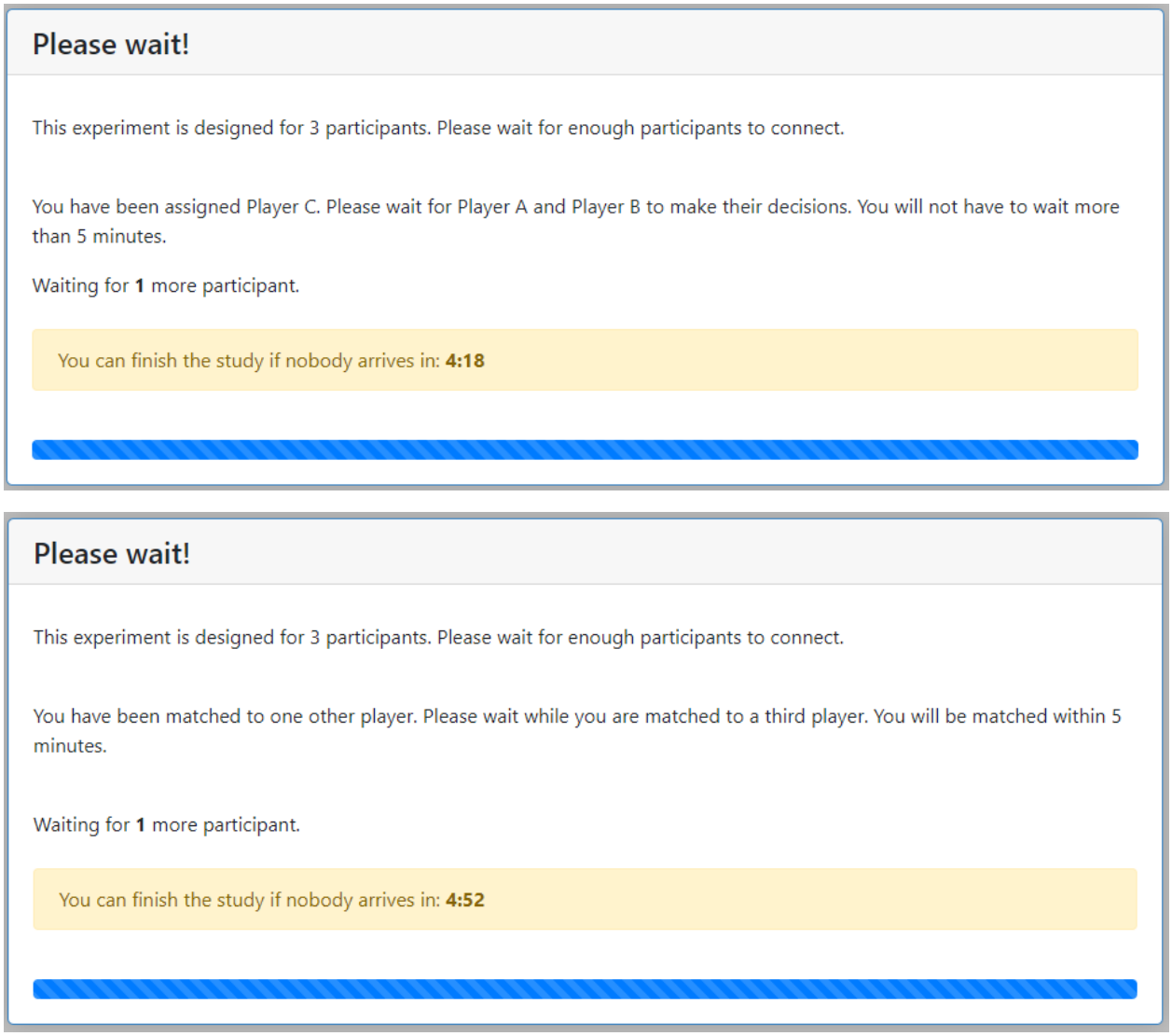
4. If A is going to choose "Right," which action will give B a higher profit?

- (•) Exit
- ( ) Enter

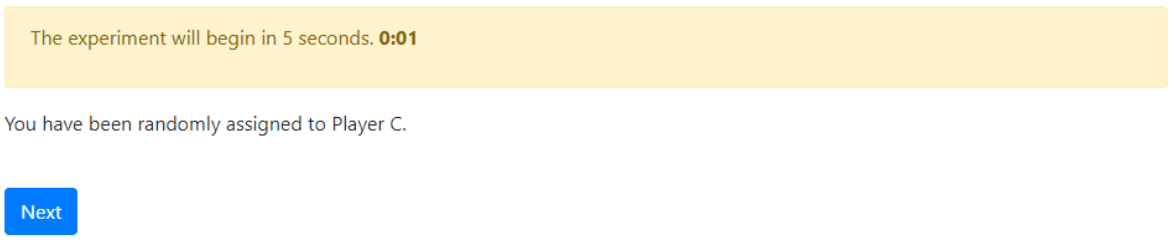
Next

Review Instructions

Players Grouping Wait page:



Start of Game:



The experiment will begin in 5 seconds. **0:02**

You have been randomly assigned to Player A.

Next

The experiment will begin in 5 seconds. **0:04**

You have been randomly assigned to Player B.

Next

Promise Screen (Player A):

Below you can choose whether or not to send a message to Player B. Both players B and C will see the message that you send.

Your Message:

- (•) I promise to choose Left.
- ( ) Do not send a message.

Select your message and then hit the button.

Next

Review Instructions

Wait Page (Player B):

Player B

Player A is making a decision. It will be your turn within 60 seconds.

Revel A Message (Player B):

Player A has not sent you a message.

You can now choose "Exit" or "Enter".

Please make your decision:

- ○ Exit
- ◉ Enter

Next

Review Instructions

Investment Page (Player B):

You have chosen to Enter.

You can now choose 0, 1, 2, or 3. Remember, that both your choice and Player A's decision will affect your final payoff.

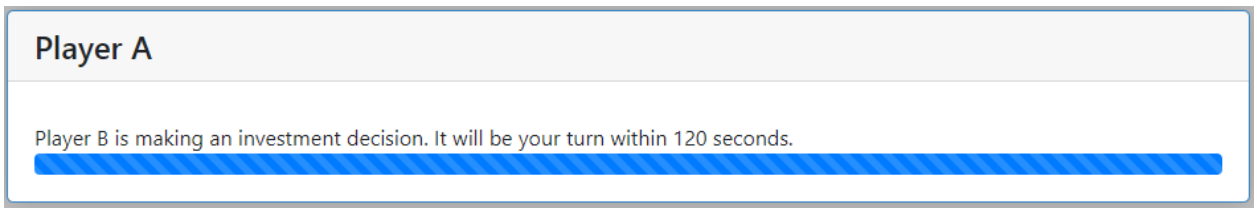
Please make your choice: :

2 ▼

Next

Review Instructions

Wait page (Player A):



Cooperation Decision (Player A):

Player B has chosen 2

You can now decide between "Right" and "Left". Remember, your payoff will depend on your choice and whether C spends any tokens reducing your payoff. Please make your decision.

Your decision?

- ○ Left
- ◉ Right

Next

Review Instructions

Reduction Slide A (Player C):

You can now decide how much to spend on reducing A's payoff. Each token that you spend will reduce your payoff by 1 token and A's payoff by 3 tokens.

If Player A promised to choose Left, Player B chose 2, and Player A chose Right, how much would you like to spend reducing A's payoff?

Choose how many tokens you would like to spend on reducing A's payoff:

3

Next

Review Instructions

Reduction Slide B (Player C):

The table below illustrates all of the possible scenarios in which A may have chosen **Right**. For each scenario, **please indicate how much you would like to spend on reducing A's payoff**. Your choice will be matched with the actual choices that A made in response to B's choice to determine the actual reduction of A's payoff.

Each token that you spend will reduce your payoff by 1 token and A's payoff by 3 tokens.

Player A promised to choose Left and A chose Right	Player B chose 0	4
	Player B chose 1	2
	Player B chose 2	3.00
	Player B chose 3	3
Player A did not promise and A chose Right	Player B invested 0	5
	Player B chose 1	1
	Player B chose 2	2
	Player B chose 3	3

Next

Review Instructions

Reduction Slide C (Player C):

The table below illustrates all of the possible choices made by A and B for which you can reduce A's payoffs, given that A **chose Left**. For each option, please indicate how much you would like to spend on reducing A's payoff. Your choice will be matched with the actual choices made by A and B to determine the punishment.

Player A **promised to choose Left** and A **chose Left**	Player B chose 0	3.66
	Player B chose 1	2.55
	Player B chose 2	5.00
	Player B chose 3	2.6
Player A **Player A did not promise** and A **chose Left**	Player B chose 0	1.8
	Player B chose 1	2
	Player B chose 2	4.56
	Player B chose 3	2.87

Next

Review Instructions

Player A Norms Elicitation (A):

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

please indicate whether you believe that A's choice of "Right" would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate.

We will determine which response was selected by the most participants assigned the role of Player A. If you give the response that is most frequently given by other people, you will receive a bonus of $0.10.

For the scenario below indicate how appropriate or inappropriate it would be for Player A to choose "Right."

Scenario	Very Inappropriate	Moderately Inappropriate	Somewhat Inappropriate	Appropriate
A made a promise to choose Left and B chose 2	◉	○	○	○

Next

Review Instructions

Player A Norms Elicitation (B):

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

The tables below give a list of possible scenarios. For each of the scenarios, please indicate whether you believe that A's choice of "Right" would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate.

For each of the choices, we will determine which response was selected by the most participants assigned the role of Player A. If you give the response that is most frequently given by other people, you will receive a bonus of $0.10 for each correct answer.

For the scenarios below indicate how appropriate or inappropriate it would be for Player A to choose "Right."

Scenario	Very Inappropriate	Moderately Inappropriate	Somewhat Inappropriate	Appropriate
Player A did not make a promise and Player B chose 0	○	○	○	◉
Player A did not make a promise and Player B chose 1	○	○	◉	○
Player A did not make a promise and Player B chose 2	◉	○	○	○
Player A did not make a promise and Player B chose 3	○	○	◉	○

Player A's Choice	Very Inappropriate	Moderately Inappropriate	Somewhat Inappropriate	Appropriate
A made a promise to choose Left and B chose 0	○	◉	○	○
A made a promise to choose Left and B chose 1	○	○	○	◉
A made a promise to choose Left and B chose 2	◉	○	○	○
A made a promise to choose Left and B chose 3	○	○	○	◉

Next

Review Instructions

Player B Norms Elicitation (1):

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

For the scenario below, please indicate whether you believe that A's choice of "Right" would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate.

We will determine which response was selected by the most participants assigned the role of Player B. If you give the response that is most frequently given by other people, you will receive a bonus of $0.10.

For the scenario below indicate how appropriate or inappropriate it would be for Player A to choose "Right."

Scenario	Very Inappropriate	Moderately Inappropriate	Somewhat Inappropriate	Appropriate
A made a promise to choose Left and B chose 2	○	○	◉	○

Next

Review Instructions

PlayerB Norms Elicitation (2):

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

For the scenario below, please indicate whether you believe that A's choice of "Right" would be very inappropriate, moderately inappropriate, somewhat inappropriate or appropriate.

For each of the choices, we will determine which response was selected by the most participants assigned the role of Player B. If you give the response that is most frequently given by other people, you will receive a bonus of $0.10 for each correct answer.

For the scenarios below indicate how appropriate or inappropriate it would be for Player A to choose "Right."

Scenario	Very Inappropriate	Moderately Inappropriate	Somewhat Inappropriate	Appropriate
Player A did not make a promise and Player B chose 0	○	○	◉	○
Player A did not make a promise and Player B chose 1	○	○	○	◉
Player A did not make a promise and Player B chose 2	○	◉	○	○
Player A did not make a promise and Player B chose 3	○	○	◉	○

Player A's Choice	Very Inappropriate	Moderately Inappropriate	Somewhat Inappropriate	Appropriate
A made a promise to choose Left and B chose 0	○	○	○	◉
A made a promise to choose Left and B chose 1	○	○	○	◉
A made a promise to choose Left and B chose 2	◉	○	○	○
A made a promise to choose Left and B chose 3	○	○	◉	○

Next

Review Instructions

Elicitation of Reason (Freeform input) (Player A):

Please briefly describe what factors influenced your decision about making a promise:

Please briefly describe what factors influenced your decision about choosing Left or Right:

Next

Elicitation of Reason (Freeform input) (Player B):

Please briefly describe what factors influenced your decision to Enter or Exit:

Please briefly describe what factors influenced your decision to choose 0, 1, 2, or 3:

Next

Elicitation of Reason (Freeform input) (Player C):

Please briefly describe what factors influenced your decisions about reducing Player A's payoff:

Next

Survey Questions (1):

Survey Questions (1 of 2)

Please answer the following questions (your response to these questions will have no effect on your payout)

	Yes	No
I did not take the scenario seriously. I just wanted to earn the fee as quickly as possible	○	◉
I carefully read the instructions	◉	○
I chose my answers to make myself seem like a good person	○	◉
This is the first time that I completed this survey	◉	○

Next

Review Instructions

Survey Questions (2) (Demographics):

Survey Questions (2 of 2)

Please answer the following questions. These will have no effect on your payout.

What is your age?

35

What is your gender?

- Male
- Female
- Non-binary
- Prefer not to answer

What is your native language?

- English
- Other

What is your highest education level?

- Some high school
- Completed high school
- Some college
- Completed college
- Post-college (i.e. graduate) degree (e.g. Masters, Phd)

Are you an mTurk Master Worker?

- Yes
- No
- I don't know what an mTurk Master Worker is

Were there any technical problems that you experienced?

No

How would you rate the clarity of the instructions?

- Very unclear
- Somewhat unclear
- Somewhat clear
- Very clear

Was any part of the experiment confusing?

No

Do you have any other comments for the researchers?

Next

Review Instructions

Dropout Page

End of Study

Unfortunately you cannot proceed further in the experiment.

Please click the "Next" button to get your Exit code.

Next

Results:

Results

The results are shown in the following table.

Player A message	1
Player B decision	Exit
Player B invested	-
Player A Cooperation decision	-
Player C punishment decision	-
Your bonus for carefully reading the Instructions	$1
Your total Payoff	$3.10

Please click "Next" to get your Exit code.

Next

Review Instructions

Kicked out participants (On Detailed Instructions page):

Unfortunately you cannot further participate in the game!

Next

Exit Codes Assignments:

Checkout

You have finished the study. Thank you for your time! In order to receive your payment you must copy and paste the following code back to Amazon Mechanical Turk:

```
a8ec252f
```

Your payment will be processed typically within the next 48 hours. If you encounter problems submitting this HIT, please search for a HIT called "ETH DeSciL Trouble Ticket" and report your problem there.

After the entire session of the experiment has concluded, we will calculate the most common response, and will apply any bonuses for choosing the most common response (only applies to players A and B).