Promises and Expectations*

Florian Ederer† Yale University

Alexander Stremitzer‡ UCLA

April 16, 2014

Abstract

We investigate why people keep their promises in the absence of external enforcement mechanisms and reputational effects. In a controlled economic laboratory experiment we show that exogenous variation of second-order expectations (promisors' expectations about promisees' expectations that the promise will be kept) leads to a significant change in promisor behavior. We document for the first time that a promisor's aversion to disappoint a promisee's expectation leads her to keep her promise. We propose a simple theory of lexicographic promise keeping that is supported by our results and nests the findings of previous contributions as special cases.

Keywords: promises, expectations, beliefs, contracts

JEL Classification: A13, C91, D03, C72, D64, K12.

*We thank Mark Greenberg whose insightful comments influenced the design of our experiment. We are also grateful to Jason Abaluck, Jennifer Arlen, Timothy Armstrong, Arthur Campbell, Gary Charness, Judy Chevalier, Martin Dufwenberg, Constança Esteves-Sorenson, Craig Fox, Robert Gibbons, Phil Haile, Holger Herz, Lisa Kahn, Navin Kartik, Amanda Kowalski, Rosario Macera Parra, Mushfiq Mobarak, Barry Nalebuff, Stefan Palan, Heikki Rantakari, Frédéric Schneider, Marta Serra Garcia, Seana Shiffrin, Joel Sobel, Rebecca Stone, Shyam Sunder, Nathan Yang, Noam Yuchtman and seminar audiences at NYU, UCLA, Yale, and at the UCSD Deception, Incentives and Behavior Conference for helpful comments and suggestions as well as Estela Hopenhayn and the California Social Science Experimental Laboratory (CASSEL) at UCLA for helping us to conduct the experiments.

†Yale School of Management, 165 Whitney Avenue, New Haven, CT 06511, florian.ederer@yale.edu.

‡UCLA School of Law, 385 Charles E. Young Drive, 1242 Law Building, Los Angeles, CA 90095, stremitzer@law.ucla.edu.

1 Introduction

To facilitate production and exchange over time, parties often make promises in order to commit to a particular course of action. There are three main reasons why a party would honor such an obligation. The first is the existence of a third party enforcement mechanism.[1] A second reason stems from reputational concerns that arise when parties are concerned that reneging on a promise might hurt their future payoffs.[2] A third reason, and the focus of the present paper, is the moral force of promise keeping.[3] A string of recent studies offer experimental evidence that promises, even if they come in the form of mere cheap talk, considerably enhance subsequent levels of cooperation in experimental hold-up, trust, and dictator games (Ellingsen and Johannesson 2004**?**, Charness and Dufwenberg 2006**?**, Vanberg 2008**?**).

However, while the practical relevance and effectiveness of the moral force of promise keeping is undisputed, there is a vigorous debate in economics, social psychology, philosophy, and law about why people keep (or should keep) their promises in the absence of explicit contractual and reputational concerns.[4] A clear understanding of what drives the desire to keep one's promise is essential to harnessing the beneficial effects of promises in institutional design, whether it be in the design of legal policy, regulatory regimes, contracts, and organizations.[5] This paper is the first to provide clean experimental evidence that promisors (senders of promises) are motivated to keep their promises in order to avoid disappointing the expectations created in the promisees (receivers of their promises). We propose a simple theory of *lexicographic* promise keeping that is supported by our results and nests the

[1] This is assumed in the economic literature on formal contracts beginning with Mirrlees (1976**?**) and Holmström (1979**?**).

[2] Such self-enforcing contracts have been studied extensively in the literature on relational contracting (Macaulay 1963**?**, Klein and Leffler 1981**?**, Bull 1987**?**, Kreps 1990**?**, MacLeod and Malcomson 1989**?**, Levin 2003**?**).

[3] Note that in law every legally enforceable contract can at least partially rely on the moral force of promise keeping as a source of commitment. This is because the legal definition of a contract requires the existence of a promise as one of its elements (Restatement 2d of Contracts §1).

[4] Notable contributions to the broader literature on promise keeping in political sciences and social psychology include Ostrom, Walker, and Gardner (1992)**?**, Kerr and Kaufman-Gilliland (1994)**?**, Sally (1995)**?**, and Bicchieri and Lev-On (2007)**?**. In legal philosophy classic references include Fried (1981)**?**, Atiyah (1983)**?**, and Scanlon (1999)**?**. For a recent contribution containing a survey of the previous literature, see Shiffrin (2008)**?**.

[5] Finan and Schechter (2012)**?** show that the moral forces of promise keeping also exist outside the lab and significantly influence political behavior. Even with secret balloting (and thus unobservable votes), promises to vote for a particular candidate are sufficiently strong for vote-buying "candidates to target trustworthy voters who can be trusted to keep their promise" (p. 876).

findings of previous contributions as special cases.

Two leading explanations for the moral force of promise keeping have been proposed. Proponents of the expectation-based theory argue that promisors keep their word in order to avoid guilt when failing to meet the expectations the promise has created in promisees. A promisor is therefore more likely to keep her promise if she believes that the promisee expects her to keep her promise.[6] In contrast, the commitment-based theory claims that promisors have a preference for keeping their word independent of the expectations of promisees.[7] While previous experimental research has either failed to distinguish between these two competing explanations or has only documented incontrovertible support for the commitment-based theory, our paper finds clean experimental evidence for the expectation-based explanation of promise keeping. By exogenously varying expectations while keeping commitment constant, we document individuals' inclination to meet the expectations induced by promises.

In theory, the two accounts are not mutually exclusive. However, Vanberg (2008)**?** presents evidence from a dictator game to show that it is commitment *per se* that induces people to keep promises and that promisees' expectations do not matter. Similarly, Ellingsen et al (2010)**?** find no significant relationship between expectations and contributions in dictator/trust games in which they elicit expectations from recipients and communicate those beliefs to the dictator/trustee. These results suggest that making promisees' expectations salient in institutional design would not be an effective way to induce more promise keeping.

In contrast to these previous studies we can show that promisees' expectations play a crucial role when promisors decide whether to keep their promises in a trust game. However, this is only the case if there is a direct promissory link between the promisor and the promisee. Where such a link is missing we find that recipients' expectations do not matter. Yet, this special case is exactly the scenario considered in previous studies. In Vanberg (2008), when deciding whether to keep their promises promisors are completely certain that they no longer face the promisees they initially promised to. In Ellingsen et al (2010), no promises are ever made.[8] In light of our results it is therefore not surprising that both studies find no significant

[6] Dufwenberg and Gneezy (2000)**?**, Charness and Dufwenberg (2006)**?** and Beck et al (2013)**?** provide some experimental evidence in support of expectations-based theories.

[7] Experimental evidence for the commitment-based explanation for promise keeping can be found in the contributions of Braver (1995)**?**, Ostrom, Walker, and Gardner (1992), Ellingsen and Johannesson (2004), and Ismayilov and Potters (2012)**?**.

[8] The recipients are merely asked about their expectations and those expectations are communicated to the dictator (trustee). Yet, the dictator has no reason to believe that those expectations were somehow

relationship between expectations and contributions. However, the main contribution of our paper is to show that this result does not generalize to the case where expectations *are* supported by promises. We therefore find support for the hypothesis that promise keeping has a lexicographic structure: A promisee's expectations matter if and only if a direct promissory link exists between the promisor and the promisee.[9]

To study the effect of differential levels of promisees' expectations for a given (high and hence behaviorally relevant) level of commitment within the same promisor-promisee pair, we modify Charness and Dufwenberg's (2006) and Vanberg's (2008) experimental designs that use trust/dictator games. In our design, a dictator can make a promise to a recipient and the recipient can decide whether to trust the dictator and to remain in the game. This choice occurs before both parties learn about the probability that it will be technically possible for the dictator to keep her promise. Imagine, for example, a situation in which the dictator promises to ship goods to the recipient. After the recipient has chosen to rely on the dictator's promise, parties learn about the probability with which the canal through which the dictator's ship needs to pass is actually open. This probability can be high or low depending on the level of political tensions. In other words, it may be very likely or very unlikely that the dictator is *able* to ship the goods and perform conforming to her promise. Accordingly, the recipient's expectations that there will be performance are higher in the high probability scenario than in the low probability scenario. In the next step, the dictator—but not the recipient—learns whether the canal is open and decides whether she wants to keep her promise.

This design allows us to compare promise keeping rates among dictators who are both able to keep their promises but hold different second-order beliefs (beliefs about how much the receiver expects to receive), depending on whether the history of the game leading up to the dictator's decision reveals that it was likely ("reliable random device") or unlikely ("unreliable random device") that the dictator would be able to perform. Using a within-subject design that allows us to observe dictators under both reliability settings, we show that the exogenous variation of the random device with which the game is played directly affects the recipient's first-order and the dictator's second-order expectations, and that these

caused by behavior attributable to her, let alone by a promise.

[9] Note that common law seems to track our lexicographic account of promise keeping. Under the doctrine of promissory estoppel, in the absence of an enforceable contract a disappointed party can recover against another party only if the other party made a promise that led to detrimental reliance.

significantly change the dictator's decision to keep her promise. Our findings provide clean evidence for an expectation-based explanation of promise keeping, as the commitment created through promises between the two parties remains constant while second-order beliefs change, thus inducing changes in promisor's performance rates. Finally, using a simple structural model we recover subject-specific susceptibilities to guilt aversion and characterize the distribution in the subject population.

The remainder of the paper is organized as follows: Section 2 presents the design of the experiment, a simple model of promises and guilt aversion, and the experimental procedures. In Section 3 we report our results. Section 4 concludes. In Appendix A we outline the instructions for the subjects participating in our experiment. Appendix B contains the formal proofs for our theoretical predictions.

2 Experimental Design and Procedure

We design an experiment to investigate the role of expectations in promise keeping. We hypothesize that a dictator is more likely to keep her promise if she believes that the recipient expects her to keep her promise. The underlying rationale for our hypothesis is that a dictator cares about the recipient's expectations if and only if those expectations are supported by the dictator's own promise.

Previous experiments were not particularly well suited to investigating this question, as they either confounded expectations- and commitment-based explanations or used the expectations created by other promisors as a means of varying the level of expectations of promisees. Thus, our design is the first to shed light on what we refer to as the lexicographic structure of promise keeping. Instead of varying second-order beliefs through the promises given by a third party as in Vanberg (2008), the promissory link between the two parties is not broken in our experiment. Rather, the magnitude of the dictator's expectations is exogenously varied by the type of random device that is selected.

2.1 Experimental Design

In our experiment, subjects are randomly matched together in pairs in each period and play the experimental trust game depicted in Figure 1: after hearing a message from the dictator, the recipient decides to opt in or out of the game and the dictator subsequently decides how

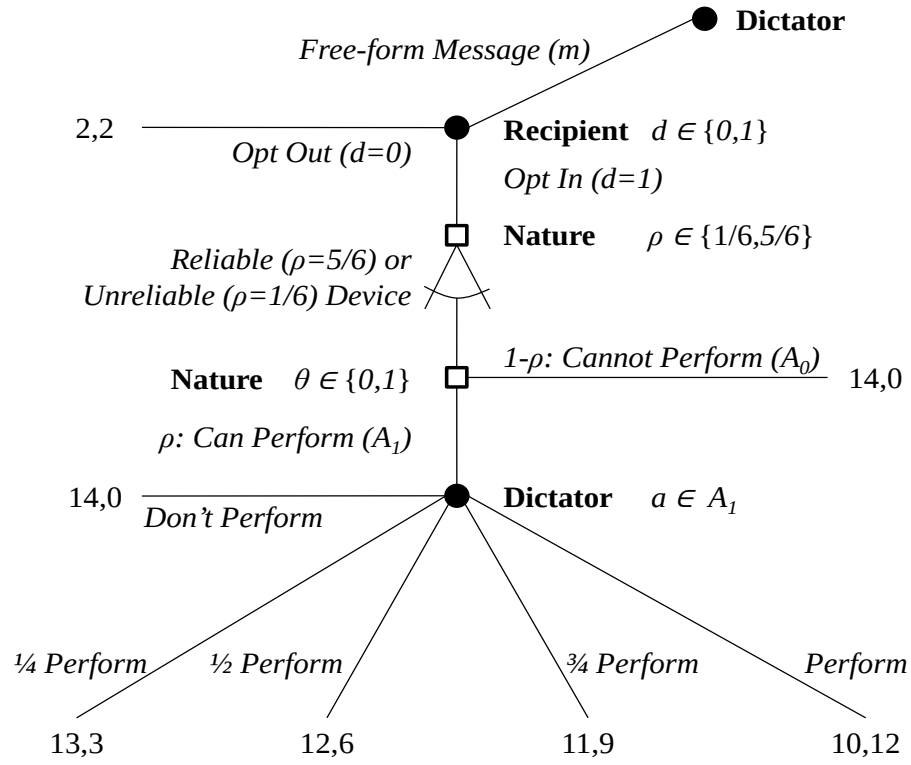


Figure 1: Dictator game with opt-out choice and reliable/unreliable device

much to contribute to the recipient.

The main feature of our design is that, following the opt-out decision of the recipient, nature selects whether the subjects play the game with a *reliable* or an *unreliable* device. This device determines how likely it is that the dictator will be able to choose some positive level of performance (i.e., any action other than *Don't Perform*). If the random device is reliable, the dictator can choose an action that delivers a positive payoff to the recipient (*Perform, 3/4 Perform, 1/2 Perform, 1/4 Perform*) with probability 5/6. If the random device is unreliable, the dictator can only deliver on a promise to perform with probability 1/6. For example, if performance is impossible, the dictator receives \$14 and the recipient receives \$0. If performance is possible and the dictator chooses *Perform*, she receives \$10 and the recipient receives \$12. Figure 1 depicts the remaining payoffs for the two players. If the dictator chooses *Don't Perform* the parties receive the same payoffs $(14, 0)$ as if performance had not been possible.

Formally speaking, the random device determines how likely it is that the dictator finds

herself in one of two states of the world, $\theta \in \{0, 1\}$, with associated action space A_θ. This action space depends on whether the dictator is able to perform, $\theta = 1$, or not $\theta = 0$:

$$a \in A_\theta = \begin{cases} A_0 = \{0\} & \text{if } \theta = 0 \\ A_1 = \{0, .25, .5, .75, 1\} & \text{if } \theta = 1 \end{cases}.$$

Figure 1 illustrates that the dictator's monetary payoffs can be written as a function of her decision a by $\pi_D(a) = 14 - 4a$. Similarly the recipient's monetary payoffs are represented by $\pi_R(a) = 12a$.

The timing of the game is as follows. At the beginning of each period ($t = 0$), subjects are randomly paired and nature randomly determines the identity of the second mover (dictator, promisor) and the first mover (recipient, promisee) in each pair. At $t = 1$ the dictator can send a free-form message m that, following the literature on promises, is coded as no message, empty talk, or promise, $\mu(m) \in \{\emptyset, 0, 1\}$.[10] Although we distinguish between the different messages in our experimental data, for the purpose of our model, we will treat no message and empty talk as the same category such that $\mu(m) \in \{0, 1\}$. At $t = 2$, the recipient decides to opt out or to stay in, $d \in \{0, 1\}$. If $d = 0$, the game ends and payoffs $(2, 2)$ for the dictator and the recipient are realized. If $d = 1$, the game continues to $t = 3$. At $t = 3$, nature determines the type of the random device ρ with which the game is played. The random device can be either *reliable* ($\rho = 5/6$) or *unreliable* ($\rho = 1/6$) where ρ denotes the probability that the dictator will be able to choose from action space A_1. Both parties learn the type of the random device. At $t = 4$, the dictator but not the recipient learns the state of the world θ and she makes the decision $a \in A_\theta$. At $t = 5$, both players learn their payoffs, and the recipient learns state θ.

In addition to recording the dictator's choice a at $t = 4$ in our experiment, we elicit the recipient's and the dictator's beliefs at $t = 3$. The recipient is asked which action $a \in A_1$ the dictator is going to choose if she will be able to perform at $t = 4$. As the recipient knows with which random device ρ the game is played, these beliefs might depend on the history of the game (realizations of μ and ρ). We therefore denote the first-order belief of the recipient by $\tau_R(\mu, \rho) \in A_1$. Furthermore, let the second-order belief $\tau_D(\mu, \rho) \in A_1$ denote the belief the dictator has about the recipient's belief $\tau_R(\mu, \rho)$, which we also elicit at $t = 3$.[11] It

[10] This freeform message approach follows previous research by Charness and Dufwenberg (2006) and Vanberg (2008). In contrast, Charness and Dufwenberg (2010)**?** use pre-coded messages and find only small effects of such 'bare' promises.

[11] As we mention below in our description of the experimental procedure, this elicitation of beliefs was incentivized.

is important to note at $t = 4$ (i.e., at the time when the dictator makes her decision) the dictator—but not the recipient—knows θ (i.e., whether the dictator can perform).

Except for the slightly richer action set for the dictator and a within-subject design, this game is to a large extent identical to the trust game in Charness and Dufwenberg (2006) and Vanberg (2008).[12] We chose a richer action space to allow for more variation in the contribution rates of dictators as well as a within-subject design that asks dictators to choose actions for both the reliable and unreliable device to increase statistical power.[13] However, what really distinguishes our design from the previous two papers is the existence of the random device, which determines the probability ρ with which the dictator will be able to choose some positive level of performance. The main purpose of this design innovation is to exogenously vary the dictator's and the recipient's expectations without breaking any promissory link that exists between a dictator and a receiver.[14]

If the random device is reliable, then there is a probability of $\rho = 5/6$ that the dictator will be able to choose *Perform*. If, on the other hand, the random device is unreliable, there is only a probability of $\rho = 1/6$ that the dictator will be able to choose *Perform*. Thus, recipients who are playing the game with an unreliable random device can plausibly expect lower monetary payoffs from the game. Because dictators are aware of this change in expectations (due to independent variation in the experiment), their second-order expectations are also exogenously changed. It is important to note that our manipulations cannot affect the commitment per se because at the time the promise is made, the dictator only knows that the game is potentially played with the reliable or the unreliable random device, but not which of the two scenarios will subsequently be realized. Similarly, at the time the dictator makes her decision at $t = 4$, she—but not the recipient—knows whether she is able to perform independent of which random device the game was played with. At this point, only the history of the game differs. If higher second-order expectations lead to higher contribution rates by the dictators who promised to perform, this would constitute evidence for the expectation-based explanation for promise keeping.

[12] Vanberg (2008) also uses euros rather than dollars, thus generating slightly higher payoffs for the participants.

[13] Brandts and Charness (2000)**?** have shown that decisions elicited using the strategy method do not differ significantly from those elicited using the "hot" method.

[14] As discussed before, in Vanberg (2008) recipient expectations are exogenously varied by using third-party promises, but this variation comes at the expense of a broken promissory link between the two parties. In Ellingsen et al (2010) no promises are ever made to begin with.

To state our ideas more formally, we now present a simple model that builds on psychological game theory (Geanakoplos, Pearce and Stacchetti 1989**?**, Battigalli and Dufwenberg 2007**?**, 2009**?**) and captures the effect of guilt aversion on promise keeping.

2.2 A Simple Model of Promises and Guilt Aversion

Battigalli and Dufwenberg (2007) propose two general theories of guilt aversion based on simple guilt and guilt from blame, respectively. Our model uses the former concept of simple guilt in which the dictator cares about the extent to which she lets the recipient down. We define $\gamma_D \geq 0$ as a constant measuring the dictator's sensitivity to guilt from disappointing the recipient's expectations, which the dictator expects to be equal to $E[\pi_R|\tau_D, \rho] = \rho 12\tau_D(\rho)$. We conjecture that promise keeping has a lexicographic structure. Guilt aversion only plays a role if the dictator sent a promise ($\mu = 1$), but not if she did not send a promise ($\mu = 0$) to the recipient.

The dictator's utility U_D when she chooses a at $t = 4$ can now be written in the following way:

$$\begin{aligned} U_D(a) &= \pi_D(a) - \frac{\mu\gamma_D}{k}\left(\max\left\{E[\pi_R|\tau_D, \rho] - \pi_R(a), 0\right\}\right)^k \\ &= 14 - 4a - 12^k\frac{\mu\gamma_D}{k}\left(\max\left\{\rho\tau_D(\rho) - a, 0\right\}\right)^k. \end{aligned} \tag{1}$$

The last term of the dictator's utility function captures the impact of guilt. This term only plays a role if the dictator sent a promise ($\mu = 1$) and if the dictator is susceptible to guilt aversion ($\gamma_D > 0$). Guilt from disappointing the recipient's perceived expectations $E[\pi_R|\tau_D, \rho]$ by choosing a low payoff $\pi_R(a)$ for the recipient has a negative effect on utility, but there is no gain from exceeding the recipient's expectations. The dictator can reduce the negative utility from guilt by increasing her action a up to the point where it matches the dictator's beliefs about the recipient's expectations. In contrast to Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2007, 2009), we allow guilt to be linear or convex in the difference between the dictator's expectations, $E[\pi_R|\tau_D, \rho]$, and the realized payoffs for the recipient, $\pi_R(a)$. For $k = 1$, our model nests the model of Charness and Dufwenberg (2006) as a special case that only admits corner solutions of a. For $k > 1$, interior solutions of a are also possible. Finally, and most importantly, note the role of the random device. When the random device is reliable ($\rho = 5/6$), the impact of guilt—by virtue of the larger expected payoff for the recipient—is much larger than when the device is unreliable ($\rho = 1/6$).

2.2.1 Predictions

It is instructive to establish two benchmark cases in which expectations do not affect actions. First, a dictator who is motivated solely by her own monetary payoff and is not sensitive to guilt at all, $\gamma_D = 0$, would have a utility of $14 - 4a$ and would therefore maximize her payoff by choosing $a = 0$.[15] Second, in the settings considered by Vanberg (2008) and Ellingsen et al (2010), in which no direct promissory link between the dictator and the recipient exists and thus $\mu = 0$, beliefs about expectations $\tau_D(\rho)$ also do not matter and hence the predictions are the same as in our first benchmark.

Thus, our model requires two assumptions for second-order beliefs to play a role in promise keeping. There must exist a promise between the two parties, $\mu = 1$, and the dictator must experience some guilt aversion, $\gamma_D > 0$. Second-order expectations will then generate different predictions about the contribution choice a for the reliable ($\rho = 5/6$) and the unreliable ($\rho = 1/6$) device. The dictator's utility is given by

$$U_D = 14 - 4a - 12^k \frac{\gamma_D}{k} \left(\max\left\{\rho\tau_D(\rho) - a, 0\right\}\right)^k \tag{2}$$

that yields different levels of guilt for the two different devices. It is straightforward to see that the impact of guilt is larger for $\rho = 5/6$ than for $\rho = 1/6$ thus leading to a higher equilibrium action a for two reasons. First, there is a difference in actions resulting purely from the exogenous variation in the reliability ρ of the device. Second, there is an additional (second-order) effect resulting from the impact of this exogenous variation on equilibrium beliefs τ_D. The first-order condition with respect to a yields the following interior solution:

$$a^* = \rho\tau_D(\rho) - \left(\frac{4}{\gamma_D 12^k}\right)^{\frac{1}{k-1}} \tag{3}$$

The dictator's action a^* is increasing in the reliability of the random device, ρ, the dictator's second-order belief, τ_D, and her susceptibility to guilt aversion, γ_D.

To see the first effect of ρ on a, assume that second-order beliefs about actions are the same in both settings, $\tau_D(5/6) = \tau_D(1/6)$, and that, just for ease of exposition, $k = 1$. As can be seen from equation (2), the guilt experienced by the dictator when choosing $a = 0$ is $2\gamma_D\tau_D$ for $\rho = 1/6$, which is much smaller in magnitude relative to the guilt experienced

[15] Of course, there are many reasons other than guilt aversion such as social preferences or norms (Rabin 1993**?**, Fehr and Schmidt 1999**?**, Bolton and Ockenfels 2000**?**, Andreoni and Bernheim 2009**?**) that would predict an equilibrium action a other than 0.

for $\rho = 5/6$ where it is $10\gamma_D\tau_D$. This argument holds a fortiori for $\tau_D(5/6) > \tau_D(1/6)$ as is evident from the first-order condition for interior solutions of a from equation (3). Thus, in equilibrium, we expect the dictator to choose higher levels of a for $\rho = 5/6$ than for $\rho = 1/6$.[16] This leads us to the main hypothesis generated by our model.

Theoretical Prediction 1 *If there is a promise ($\mu = 1$), the dictator's contribution action a is higher for the reliable than for the unreliable device. If there is no promise ($\mu = 0$), there is no difference in the dictator's contribution action.* ***(TP1)***

As a result of these different action choices, we also expect first- and second-order beliefs to differ in the two settings. In particular, because actions are higher for $\rho = 5/6$, equilibrium first-order beliefs of recipients must adjust to the different actions that the dictator chooses in the two settings. Hence, first-order beliefs are higher, $\tau_R(5/6) \geq \tau_R(1/6)$ and, as a result, equilibrium second-order beliefs must be higher too, $\tau_D(5/6) \geq \tau_D(1/6)$. This leads us to our second prediction.

Theoretical Prediction 2 *If there is a promise ($\mu = 1$), first-order and second-order beliefs are higher for the reliable device than for the unreliable device. If there is no promise ($\mu = 0$), there is no difference in first-order and second-order beliefs.* ***(TP2)***

Note that this difference in second-order beliefs provides a second reason for why actions are higher under $\rho = 5/6$ and reinforces the first effect on the dictator's action discussed in **TP1**. Finally, because the dictator's action is higher in equilibrium if there is a promise ($\mu = 1$) compared to when there is none ($\mu = 0$), the recipient's expected payoff from not opting out is higher if $\mu = 1$. Therefore, the recipient's opt-out decision and his beliefs should be responsive to the message of the dictator. This lead us to our third hypothesis.

Theoretical Prediction 3 *The recipient's opt-in rate and first-order beliefs are higher if there is a promise ($\mu = 1$) than if there is no promise ($\mu = 0$).* ***(TP3)***

Consistent with the results obtained by Charness and Dufwenberg (2006) and Vanberg (2008), we find that promises made in the communication phase significantly increase both the rate at which dictators choose an action other than *Don't Perform* and the second-order

[16] See Appendix B for a more rigorous proof of the theoretical predictions taking the possibility of corner solutions into account.

beliefs concerning the probability that they will do so. However, the goal of the present paper is to investigate whether independent variation in the level of second-order beliefs achieved through the use of a random deviceleads to different performance levels by dictators who previously made promises.

2.3 Experimental Procedure

We conducted 20 experimental sessions with a total of 280 student subjects at the California Social Science Experimental Laboratory (CASSEL). The CASSEL subject pool consists of undergraduate students from UCLA. Subjects were assigned to visually isolated computer terminals. Beside each terminal they found paper instructions, which are reproduced in Appendix A. Questions were answered individually at the subjects' seats.

Each session consisted of 2 unpaid practice rounds followed by 8 paying rounds. In each round, subjects interacted with another randomly chosen participant. Under no circumstances did any participant interact with the same participant twice in the paying rounds. We achieved this by creating matching groups of exactly 10 subjects. At the end of the experiment, one of the 8 paying rounds was randomly chosen for payment. Each round was equally likely to be selected. The amount paid out at the end of the experiment depended on the decisions made in that round. In each period we also elicited first- and second-order beliefs of subjects about the behavior of other subjects. This elicitation of beliefs was incentivized and subjects were paid for all rounds except the one chosen for payment of the decision to prevent hedging. The subjects received a fixed fee of $10 for arriving on time. The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).**?**

First, each subject was randomly matched with an interaction partner, and one participant in each pair was randomly assigned to the role of a dictator or the role of a recipient.[17] The pair matches and the roles of the players were randomly assigned anew in each round. It was always equally likely to be assigned to either role, regardless of the previous messages or actions in the game.

Second, each dictator could send free-form messages to the recipient. The dictator could send any number of (unidirectional) messages of a length of 256 characters each within a

[17] In the instructions, we neutrally refer to the role of the dictator and the role of the recipient as "Role A" and "Role B," respectively.

time frame of 90 seconds.[18] As in Vanberg (2008), subjects were not allowed to reveal their identity to the other participant. That is, they were not allowed to reveal their name or any other identifying feature such as race, gender, hair color, or seat number. In every other respect, subjects were free to send any message they liked.

Third, after receiving the message of the dictators, the recipients could decide whether to opt out. If a recipient chose to opt out, each player received \$2. If a recipient chose not to opt out, the game continued. At this point, neither player knew whether the random device determining whether the dictator would be able to perform was *Random Device 5/6* ("reliable random device", probability of 5/6 that dictator would be able to choose something other than *Don't Perform*) or *Random Device 1/6* ("unreliable random device", probability of 1/6 that dictator would be able to choose something other than *Don't Perform*). However, both parties knew that each scenario could occur with an equal probability of 50%.

Fourth, nature privately determined whether the players would play the game with the reliable or the unreliable random device, which determined the probability with which the dictator in the pair was able to choose a positive level of performance. At this point neither player learned with which type of random device the game was played. The players only learned at the end of each round which random device was chosen.

Fifth, the recipient guessed which choice the dictator would likely make if she could choose to perform and the dictator guessed which payoff the recipient expected to earn. Specifically, recipients and dictators were asked to choose from a five point scale. While the recipient's guessing payoff depended on the contribution decision of the dictator, the dictator's guessing payoff depended on the belief chosen by the recipient. Both payoffs were higher the closer they were to the actual contribution and belief, respectively.[19] As both parties did not know which random device had been chosen, we asked the players to make their guesses for both reliability scenarios. Note that if a recipient thought that the dictator intended to choose *Perform* (allocating \$12) a recipient's expected payoff depended on the reliability scenario. The expected payoff was \$2 if the game was played with Random Device 1/6 (12*1/6=2) and \$10 if the game was played with Random Device 5/6 (12*5/6=10). Asking the dictator to make her guesses in terms of the recipient's expected payoffs allowed

[18] Note that the 90-second time frame was not enforced, as it just served as an informal pacemaker.

[19] With the exception of five (rather than two) potential choices for the contribution decision our belief elicitation method is identical to that used in Vanberg (2008).

us to make those expectations particularly salient. These procedures yielded five point scales between 0 (performance very unlikely) and 1 (performance very likely) for first- and second-order beliefs.

Sixth, the dictator was asked to assume that she were able to perform and to make their contribution decisions as if the game leading up to that point had been played with Random Device 5/6 or Random Device 1/6. Figure 1 depicts the players' payoffs under the different possible contribution decisions.

Seventh, the computer randomly drew an equally likely integer between 1 and 6 for each pair using z-Tree's random number generator. If the random device was reliable it was possible for the dictator to perform for the numbers 2, 3, 4, 5, and 6. If the random device was unreliable, the dictator was able to perform for the number 1.

Finally, at the end of each round, both dictators and recipients learned with which random device the game had been played, whether it had been possible for the dictators to perform, and the payoffs both participants had earned.

3 Results

The data comprise 20 experimental sessions involving a total of 280 subjects with a total of 28 matching groups of 10 subjects. Each session lasted for 8 rounds. The average number of dictator decisions made by each subject is 4. As we used the strategy method to elicit first- and second-order beliefs and contribution choices for both types of the random device, this within-subject design gives us a total of 1,120 decisions made under each reliability scenario. However, each matching group constitutes only one independent observation. Non-parametric tests are therefore based on matching group averages of the relevant variables. For comparisons between the random devices, we have matched observations that allow us to use Wilcoxon signed-rank tests while for unmatched comparisons between message categories we use Mann-Whitney (MW) rank-sum tests.

3.1 Performance Rates

To investigate the role of promises, we asked a student assistant to code messages according to whether they contained a promise stating that the subject would choose any action other than *Don't Perform*. Following Charness and Dufwenberg (2006), this classification yielded three

categories: "no message" ($\mu = \emptyset$) containing no text at all; "empty talk" ($\mu = 0$) messages (e.g., "Hey I just met you/and this is crazy/but here's my message/so money maybe?"); and "promise" ($\mu = 1$) messages (e.g., "im going to choose 3/4th perform so please dont opt out"). After accounting for all opt-out decisions (see Section 3.4), there remain 383, 300, and 268 individual observations, and 28, 27, and 28 matching group observations in the three message categories (promise, empty talk, no message) for both the reliable and the unreliable device. We use these for our remaining analysis.

When the dictator made a promise ($\mu = 1$), the average contribution (performance rate) she gave to the recipient was \$7.08 (0.59) for the reliable ($\rho = 5/6$) and \$6.48 (0.54) for the unreliable device ($\rho = 1/6$), conditional on performance being feasible. While this difference is small in magnitude it is statistically significant (Wilcoxon signed-rank, p-value < 0.01) and consistent with our main hypothesis **TP1**. Thus, when the dictator made a promise, she actually chose higher performance rates when it subsequently turned out that it was likely that she would be able to perform (reliable device), as opposed to when the possibility of performance was unlikely (unreliable device). Furthermore, as we show below the small magnitude of the contribution differentials is in line with the small, but significant differentials in second-order beliefs between the two random device settings, lending further empirical support to the expectations-based explanation for promise keeping. Dictators contributed significantly more under the reliable device scenario because they (correctly) held exogenously higher second-order beliefs about the recipients' expectations in that scenario.

As we are employing the strategy method (and thus a within-subject design), it is particularly instructive to examine the behavior of those dictators who made different contribution decisions in the unreliable and the reliable device setting. Figure 2 shows the contribution decisions of dictators who promised to contribute and who chose to alter their contribution decision depending on whether the device was reliable or unreliable. A much lower proportion of dictators chose *Don't Perform* for the reliable than for the unreliable device, and hence more of them ended up choosing higher performance rates. While 40% of dictators chose *Don't Perform* under the unreliable device, only less than 5% chose the same action under the reliable device. As a result, a much larger proportion of dictators chose to contribute a positive amount under the reliable device, with the *3/4 Perform* action experiencing the largest increase.[20]

[20] Figure 2 also shows that many subjects make interior choices of a, suggesting that guilt aversion is not

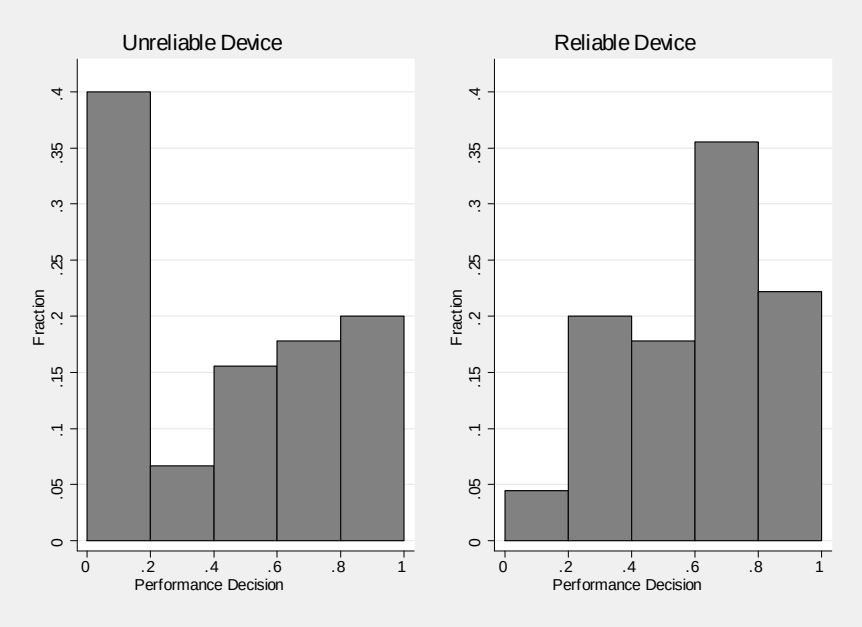


Figure 2: Fraction of performance decisions of dictators (*Don't Perform*, *1/4 Perform*, *1/2 Perform*, *3/4 Perform*, *Perform*) who sent a promise and changed their decision between the unreliable and the reliable device

At this point, it is crucial to note that the significant difference in dictator contributions between the two settings disappeared when the dictator engaged in empty talk (Wilcoxon signed-rank, p-value 0.27) or sent no message at all (p-value 0.14). Thus, in the absence of an explicit promise, higher reliability did not lead to higher performance. This finding is the empirical equivalent of the second part of **TP1**, which shows that contributions should not differ between the two reliability settings if the dictator did not promise. This empirical result is particularly important because it provides additional support for the conjectured lexicographic structure of promise keeping. The receiver's expectations influence the dictator's contribution decisions only when the dictator committed herself to a promise; they play no role when the dictator made no promise.

linear ($k = 1$), but that it is instead convex ($k > 1$) in the difference between beliefs and actions. We explore the magnitude of guilt aversion in greater detail in Section 3.5.

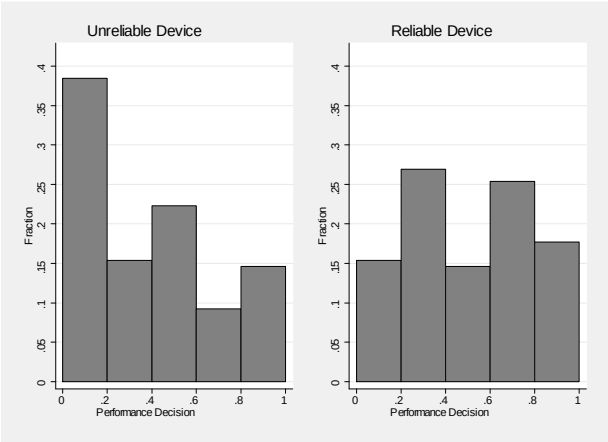


Figure 3a: Empty Talk

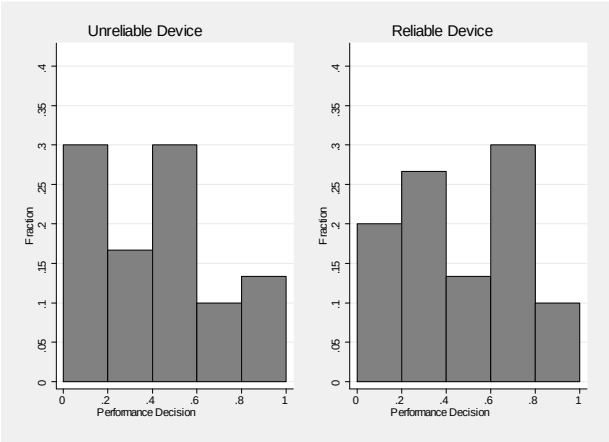


Figure 3b: No Message

The two panels of Figure 3 depict the same data as Figure 2 for those dictators who changed their contribution decision between the two settings, but instead focuses on the message categories of empty talk and no message. While there is a small change in behavior towards more generous contribution rates from the unreliable to the reliable device for dictators who sent an empty talk message (Figure 3a), there is practically no change for dictators who sent no message (Figure 3b). The slightly positive (but statistically insignificant) shift in dictator contribution rates is larger for empty talk messages than for no messages. In a related experiment that investigates commitment-based explanations of promise keeping Ismayilov and Potters (2012) find that both trustees who make a promise *and* those who do not are more likely to be trustworthy if their message is delivered to the trustor. Their findings as well as ours suggest that any form of communication increases trustworthiness irrespective of the content of messages. This tentatively suggests that the guilt aversion effect is larger when there is some communication rather than none at all. However, as pointed out above, in both cases and in contrast to when the dictator made a promise, there is no significant difference in the contribution decisions.

Finally, note that the highly statistical difference in performance rates between the reliable and unreliable device for promises (Wilcoxon signed-rank, p-value < 0.01) and the lack of statistical significance for empty talk (p-value 0.27) and no message (p-value 0.14) are not caused by large differences in the number of individual observations (and hence statistical power) across the three categories. As mentioned above, there are 383 (promise), 300 (empty talk), and 268 (no message) individual observations, and 28, 27, and 28 matching group ob-

servations in the three message categories for both the reliable and the unreliable device. In accordance with other previous contributions, we also find that average performance rates are higher if the dictator made a promise (0.56) than if there was just empty talk (0.38) or no message (0.38) (MW rank-sum, p-values < 0.01).

3.2 First-Order Beliefs and Expectations

Recipients were asked to guess the dictator's decision on a five-point scale between 0 and 1. When the random device was reliable, the recipients had mean first-order beliefs $\tau_R(\mu, 5/6)$ of 0.23, 0.26, and 0.34 for no message, empty talk and promise, respectively. These beliefs were slightly lower (0.22, 0.21, and 0.29) for the same messages in the case of an unreliable device, $\tau_R(\mu, 1/6)$. For both outcome realizations of the random device, receiving a promise (relative to receiving no message or empty talk) significantly raised the recipients' expectations about how much they would receive from the dictator, moving first-order expectations from the lower values of 0.23 and 0.26 to 0.34 and from 0.22 and 0.21 to 0.29, respectively (MW rank-sum, p-values 0.007, 0.005, 0.022, 0.006). This pattern mirrors the results of Charness and Dufwenberg (2006) and Vanberg (2008) and is consistent with **TP3** of our model.

However, although the first-order beliefs are lower across the board for the unreliable device, the differences in first-order beliefs between the two device settings are not statistically significant (Wilcoxon signed-rank, p-values 0.17, 0.95, 0.45). Even when focusing on the case in which the dictator sent a promise, the difference in the recipient's first-order beliefs is not statistically significant between a reliable and an unreliable device. The lack of a statistically significant difference in first-order beliefs between the two reliability settings is consistent with **TP2** if the dictator did not promise (no message and empty talk). However, it is inconsistent with the same theoretical prediction for the case where the dictator made a promise, because the model predicts a difference in first-order beliefs in that case.

It is important to remember that we elicited conditional beliefs that directly match $\tau_R(\mu, \rho)$ in our model and allow for easy comparability across the two reliability devices. Recipients were asked how much (on a five-point scale) they thought the dictator would contribute if performance was feasible, as the recipients did not actually learn whether performance was feasible for the dictator until after the end of each round. Therefore, the relevant first-order *expectations* at the time the dictator forms her second-order expectations and makes her performance decision are given by the unconditional first-order expectations,

which are $12\tau_R(\rho)\rho$. Thus, in order to obtain the first-order expectations in terms of expected payoffs, the elicited conditional first-order beliefs have to be multiplied by 5/6*12=10 for the reliable and by 1/6*12=2 for the unreliable device. We find that these unconditional first-order beliefs are substantially higher in the reliable ($2.60, $2.30, and $3.40) than in the unreliable scenario ($0.42, $0.44, and $0.58), and that these differences are statistically significant (Wilcoxon signed-rank, p-values < 0.01).[21]

3.3 Second-Order Beliefs and Expectations

We next investigate how second-order beliefs $\tau_D(\mu, \rho)$ (i.e., a dictator's belief $\tau_D(\mu, \rho)$ about the belief $\tau_R(\mu, \rho)$ that the recipient has about the dictator's performance decision) vary with the message sent by the dictator and the reliability of the random device. When the random device was reliable, second-order beliefs were 0.59 when the dictator sent no message, 0.59 for an empty talk message, and 0.72 when a promise was given. In contrast, when the random device was unreliable, the same second-order beliefs fell to 0.48, 0.43, and 0.64, respectively. The second-order beliefs in both reliability settings are significantly higher for promises than for empty talk and no messages (MW rank-sum, p-values < 0.01).

More importantly, though, given the focus of this paper, if the dictator made a promise the difference in second-order beliefs between the reliable and the unreliable random device is statistically significant (Wilcoxon signed-rank, p-value < 0.01) which is consistent with **TP2**. That is to say, when a dictator made a promise to share money with the recipient, her belief about the amount the recipient expected she would contribute was significantly higher when the random device was reliable than when it was unreliable. The likely reason for this finding is the dictator's realization that the recipient expected a higher level of performance when the random device was reliable than when it was unreliable and the dictator adapted her second-order beliefs accordingly.

When the dictator did not send any message or sent an empty talk message, second-order beliefs were also significantly higher in the 5/6 random device case than in the 1/6 random device condition (Wilcoxon signed-rank, p-values < 0.01). This contradicts **TP2** which predicts that second-order beliefs should not differ between the two reliability settings

[21] This stark difference is a feature of our experimental design. In contrast to previous contributions, we do not vary second-order expectations through the endogenous variation of first-order beliefs. Instead, we directly and exogenously change second-order expectations through the different random device scenarios and the timing of when the dictator and the recipient learn about which random device was chosen.

if the dictator did not make a promise. However, as shown in Section 3.1 and as predicted by our main hypothesis **TP1** contribution rates were not significantly higher if expectations were not supported by a promise.[22]

The second-order expectations given by $E[\pi_R|\tau_D, \rho] = 12\tau_D(\rho)\rho$ that correspond to these second-order beliefs influence the level of guilt experienced by the dictator subjects as can be seen from the utility function in equation (1). These second-order expectations are equal to \$5.88, \$5.91, and \$7.18 for the reliable device and significantly lower (Wilcoxon signed-rank, p-values < 0.01) for the unreliable device where they are equal to \$0.96, \$0.87, and \$1.28. This large difference in second-order expectations is, of course, mainly exogenously created by the use of the random device and this large variation serves us in our goal to cleanly identify the effect of guilt aversion on promise keeping.

3.4 Promises and Opt-Out Decisions

Recipients chose to opt out at different rates depending on which message they received from the dictator with whom they were paired. While only 7.3% of recipients chose to opt out after receiving a message classified as a promise, 21.6% opted out if they received no message at all, and 12.8% opted out if they received an empty talk message. These differential opt-out rates are consistent with **TP3**, which predicts that the recipient's (expected) payoff from staying in the game is higher if the dictator sent a promise than if she did not.

The differences in opt-out rates between recipients who received a promise and those who received an empty talk message, and between empty talk messages and no messages are only significant at the 10% level (MW rank-sum, p-values 0.07, 0.07). In contrast, the difference in opt-out rates between participants who received a promise and those who did not receive any message is significant (MW rank-sum, p-value < 0.01). Similarly, the difference between participants who received a promise and (pooled) participants who did not receive a promise, either because they received an empty talk message or no message at all, is significant at the 1% level (MW rank-sum, p-value < 0.01). The low opt-out rate of recipients who received a promise from their partnered dictator indicates that the recipients expected higher relative

[22] Dictators may have thought that even if they gave no promise recipients may hold higher (out-of-equilibrium) first-order expectations if the random device is reliable than if it is unreliable. Although this contradicts **TP2**, this out-of-equilibrium phenomenon may lend additional support to our theory of lexicographic promise keeping. Although the second-order beliefs are higher they do not translate into higher contribution rates if there is not promise. In other words, differences in second-order beliefs only matter if they are justified by a promise.

payoffs from staying in the game than from opting out compared to recipients who received no message at all or just an empty talk message. Furthermore, the higher opt-out rate for recipients who received no message relative to recipients of an empty talk message suggests that some form of verbal engagement is better than none at all when it comes to inducing recipients to stay in the game.[23]

Finally, although our experiment was specifically designed to test whether or not guilt aversion influences promise keeping, by using the variation in promises, second-order expectations and contribution decisions we can also assess the relative importance of the expectation- and the commitment-based explanations of promise keeping. However, this back-of-the-envelope calculation has to be interpreted with a certain degree of skepticism because in our experiment we do not have exogenous variation in promises to directly test the commitment-based explanation of promised keeping. We therefore estimate a linear regression model with subject-specific fixed effects in which the dictator's contribution decision is the dependent variable and the (binary) promise variable and second-order expectations interacted with the promise variable are the independent variables. From this estimation we obtain the ratio of explained variation of the two independent variables $var\left(X_1\hat{\beta}_1\right)/var\left(X_2\hat{\beta}_2\right)$ where $X_1 = \mu = \{0,1\}$ is the binary promise variable, $X_2 = \rho\tau_D(\rho)$ is the second-order expectation and $\hat{\beta}_1$ and $\hat{\beta}_2$ are the associated estimated coefficients. The specific value of $var\left(X_1\hat{\beta}_1\right)/var\left(X_2\hat{\beta}_2\right)$ is 0.513 suggesting that the variation in 2nd order expectations actually explains almost twice as much of the variation in contribution decision as variation in promises. Thus, in our particular experimental context the expectation-based explanation plays a much larger role than the commitment-based explanation of promise keeping.

3.5 Distribution of γ_D

Using our simple model of guilt aversion as well the experimental data obtained on performance actions a, beliefs τ_R and τ_D, and message categories μ, we can directly recover each dictator's susceptibility to guilt aversion, γ_D. However, because many dictators choose strictly positive levels of performance a even when there is no promise ($\mu = 0$), we augment our previous model by an additional term that captures altruism. The dictator's utility

[23] This is in line with the aforementioned findings of Potters and Ismayilov (2012) who also find that even some form of communication like empty talk increases trustworthiness.

function is then given by

$$U_D = \pi_D(a) - \mu\frac{\gamma_D}{k}\left(\max\left\{E\left[\pi_R|\tau_D,\rho\right] - \pi_R(a), 0\right\}\right)^k - \frac{\delta_D}{r}\left(12 - \pi_R(a)\right)^r$$

where the dictator suffers a convex disutility if the receiver's payoff $\pi_R(a)$ falls short of his maximum possible payoff of 12. If $\delta_D = 0$, the dictator is not driven by altruism, but as δ_D increases she cares more about the payoff obtained by the receiver. The values of k and r influence the convexity of the guilt aversion and the altruism terms, and are assumed to be known. Using identifying variation for subjects who are observed in the data choosing a under both a promise ($\mu = 1$) and no promise ($\mu = 0$), the unknown altruism and guilt aversion parameters, δ_D and γ_D, are exactly identified from the first-order conditions with respect to a. To see this ignore, just for expositional purposes, corner solutions and note that for $\mu = 0$ we have

$$0 = -4 + 12\delta_D\left(12 - 12a\right)^{r-1}$$

which allows us to obtain δ_D. Similarly, for $\mu = 1$ we have

$$0 = -4 + 12\gamma_D\left(\max\left\{12\rho\tau_D - 12a, 0\right\}\right)^{k-1} + 12\delta_D\left(12 - 12a\right)^{r-1}$$

which using δ_D yields γ_D.[24]

The two panels of Figure 4 show the distribution of the altruism and guilt aversion parameters in the dictator population for quadratic altruism, $r = 2$, and quadratic guilt aversion, $k = 2$. Given our assumptions, we are only able to identify the distribution of γ_D for 104 dictators who are observed under both $\mu = \{0, \emptyset\}$ *and* $\mu = 1$. As suggested by our reduced form analysis that documents a significant positive shift in performance from unreliable to reliable device when $\mu = 1$, the distribution of γ_D (Figure 4b) shows that more than half of the dictators exhibit guilt aversion, $\gamma_D > 0$, while the remaining, slightly smaller proportion of just under 50% is unaffected by this behavioral trait, $\gamma_D = 0$.

The dictators with a positive γ_D fall into two broad categories as can be seen from Figure 4b. First, there is a mass of about 25% of all dictators (at $\gamma_D \approx 20$) where γ_D is so large that in equilibrium the dictators raise a sufficiently high that $a \approx \rho\tau_D$. In this way, they reduce their own monetary payoff in order to completely avoid any loss from guilt aversion

[24] In our experimental setting dictators are constrained to choose $a \in \{0, .25, .5, .75, 1\}$. However, since we also have several observations of a for each dictator we use dictator-specific averages of a for given μ and ρ when solving the first-order conditions for δ_D and γ_D. We take corner solutions at $a = \{0, 1\}$ into account.

which they would otherwise suffer if they chose a lower performance a. Of course, the true γ_D of these dictators might be even higher than 20, but these subjects are already at a corner solution in our data. Second, for roughly 25% of dictators, γ_D lies between the two mass points of 0 and 20, and so these dictators trade off some monetary gains against losses from guilt aversion. They do not, however, raise a high enough to completely eliminate guilt in equilibrium.

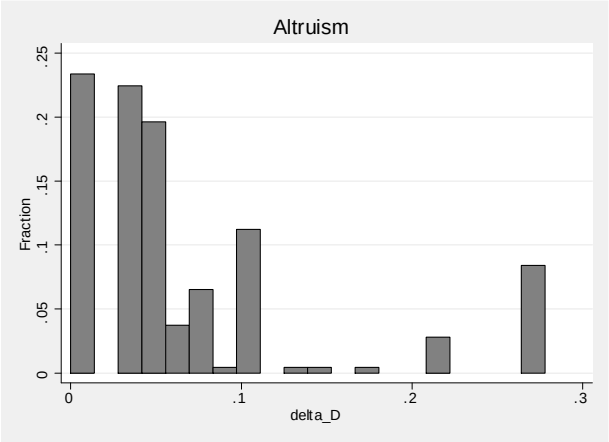


Figure 4a: Altruism parameter δ_D for $r = k = 2$

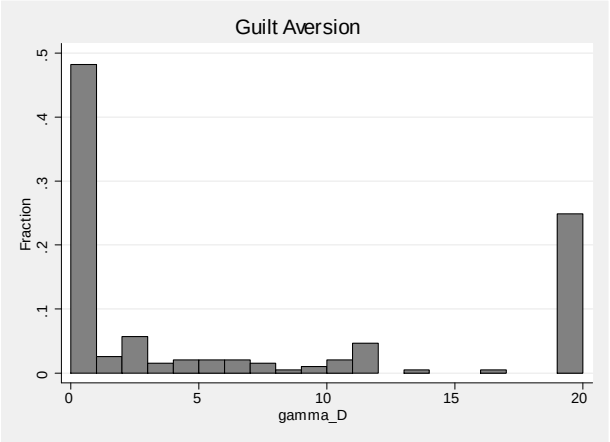


Figure 4b: Guilt aversion parameter γ_D for $r = k = 2$

4 Conclusions

Many psychological and economic experiments have shown that promises greatly enhance cooperative behavior in experimental games, but evidence on the driving forces behind why people keep their promises remains scant. In this paper we provided the first incontrovertible evidence for the expectation-based explanation of promise keeping.

Previous experiments either could not distinguish between commitment-based and expectation-based explanations, because treatment-induced changes in the alternative causal factors (promises and second-order beliefs) had occurred simultaneously, or focus on settings in which there was no sufficient level of commitment between the dictator and the recipient for the changes in the levels of (first- and second-order) expectations to affect behavior.

In contrast, we designed our experiment to achieve independent variation in second-order expectations in an environment where these were supported by a direct promissory link between the dictator and the recipient, and thus by the existence of a sufficiently high

level of commitment. Changes in the probability with which a dictator would be able to contribute directly impacted recipients' first-order and dictators' second-order expectations, which in turn significantly changed behavior. In light of our own as well as previous findings, we propose a lexicographic theory of promise keeping. Guilt and expectations matter if and only if they are supported by the promise between the two acting parties.

A Instructions

Thank you for participating in this experiment. The purpose of this experiment is to study how people make decisions in a particular situation. In case you should have questions at any time, please raise your hand. Please do not speak to other participants during the experiment. You will receive \$10 for arriving on time. Depending on the decisions made and the decisions of other participants, you may receive an additional amount (as described below). At the end of the experiment, the entire amount will be paid to you individually and privately in cash.

This session consists of 2 practice rounds and 8 paying rounds with money prizes. In each round, you will interact with another randomly chosen participant. Under no circumstances will you interact with the same participant twice. No participant will learn the identity of the persons with whom he or she has interacted during the experiment.

At the end of the experiment, one of the 8 paying rounds will be randomly chosen for payment (every round is equally likely). The amount that you will receive at the end of the experiment will depend on the decisions made in that round.

Each round consists of 7 steps, which are described below.

Overview

There are two players; Player A and Player B. Initially, A can send a chat message to B over the computer, and B can decide whether he wants to opt out of the game, leading to payoffs of \$2 for each player. If B does not opt out, a random device will determine whether it will be possible for A to perform, that is, allocate money to B. If it is impossible to perform, Player A gets a payoff of \$14 and B gets a payoff of 0. If it is possible for A to perform, he can make one of 5 choices:

- *Don't Perform*: A keeps \$14 for himself and allocates \$0 to B.
- *1/4 Perform*: A keeps \$13 for himself and allocates \$3 to B.
- *1/2 Perform*: A keeps \$12 for himself and allocates \$6 to B.
- *3/4 Perform*: A keeps \$11 for himself and allocates \$9 to B.
- *Perform*: A keeps \$10 for himself and allocates \$12 to B.

There are two types of Random Device

- Random Device 5/6: A is able to choose something other than *Don't Perform* with probability 5/6.
- Random Device 1/6: A is able to choose something other than *Don't Perform* with probability 1/6.

The players learn about the nature of the random device after B has made his opt-out decision.

Step 1: Role assignment. At the beginning of each round, you will be anonymously and randomly matched with another participant. Each member of the pair will then be randomly assigned Role A or Role B with equal probability (50%).

Step 2: Communication. During the communication phase, Player A can send a chat message to Player B. Important: You are not allowed to reveal your identity to the other participant. (That is, you may not reveal your name or any other identifying feature such as race, gender, hair color, or seat number.) In every other respect, you are free to send any message you like. Please continue to remain quiet while communicating with the other participant. Participants who violate these rules (experimenter discretion) will be excluded from the experiment and all payments.

Step 3: Opt-out decision. Player B can decide whether to opt out. If B chooses to opt out, each player receives \$2. If B chooses not to opt out, the game continues. Information: Neither player knows, whether the Random Device determining if A will be able to choose *Perform* is Random Device 5/6 (probability that A can choose something other than *Don't Perform* is 5/6) or Random Device 1/6 (probability that A can choose something other than *Don't Perform* is 1/6). However, both parties know that each scenario occurs with equal probability (50%).

Step 4: Nature of the Random Device revealed. The players learn whether they play with Random Device 5/6 or Random Device 1/6.

Step 5: Guessing. Player B guesses which choice Player A is likely to make in Step 7. A guesses which payoff B expects to gain. Note that if B thinks that A intends to choose *Perform*—allocating \$12—B's expected payoff depends on what B has learned about the Random Device: The expected payoff is \$2 if the game is played with Random Device 1/6 (12*1/6=2) and \$10 if the game is played with Random Device 5/6 (12*5/6=10).

Step 6: The Players learn whether A will be able to perform. If only *Don't Perform* is possible, the game ends. If A is able to perform, the game continues to Step 7.

Step 7: Decision phase. A decides whether to choose *Don't Perform* (keep \$14 and send \$0 to B), or whether to choose *Perform* (keep \$10 and send \$12 to B) or any of the options in between. The payoffs are

	A	B
A chooses *Don't Perform*	\$14	\$0
A chooses *1/4 Perform*	\$13	\$3
A chooses *1/2 Perform*	\$12	\$6
A chooses *3/4 Perform*	\$11	\$9
A chooses *Perform*	\$10	\$12
B chooses "Opt Out"	\$2	\$2
Performance not possible	\$14	0

Information at the end of a round. Players learn their own payoff, which random device was chosen, and whether A was able to perform.

Conditional Choice. You will be asked to make the guess in Step 5 and the decision in Step 7 before Step 4 has actually been played. In other words, you will be asked to assume that A will be able to perform in Step 7, and then make the guess in Step 5 and the decision in Step 7 for two scenarios:

1. Random device 1/6 was chosen.

2. Random device 5/6 was chosen.

Subsequently, Steps 4 and 6 are played and A's recorded choice will be entered as A's decision in Step 7 (provided the game reaches this step). A's decision will influence payoffs as if A took the same decision in Step 7.

Bonus: Guessing. At certain points, you will have the additional possibility to earn a small amount by guessing the decisions of the other participant. Guessing will be paid in every round that is not chosen for payment of the decision. You will learn more about this during the experiment.

Do you have any questions?

B Proofs

We use Battigalli and Dufwenberg's (2007) general model of simple guilt to capture guilt aversion in our model. Applying their formulation to our game and notation yields the following utility function for the dictator

$$U_D = \pi_D(a) - \gamma_D \max\left\{E\left[\pi_R|\tau_D,\rho\right] - \pi_R\left(a\right), 0\right\}.$$

To this formulation we add the lexicographic structure of promise keeping which is governed by μ in our model, and we allow for convex guilt to obtain

$$U_D = \pi_D\left(a\right) - \frac{\mu\gamma_D}{k}\left(\max\left\{E\left[\pi_R|\tau_D,\rho\right] - \pi_R\left(a\right), 0\right\}\right)^k$$

We will first prove **TP1** by showing that the dictator's equilibrium choice a^* after having given a promise is strictly increasing in ρ for sufficiently high guilt aversion and 0 otherwise. The prediction for $\mu = 0$ follows trivially from the dictator's utility function in expression (2). We distinguish two cases, $k = 1$ and $k > 1$.

Case $k = 1$. If $k = 1$, $U\left(a\right)$ is a linear function in a and $U'\left(a\right) = -4 + 12\gamma_D$. It follows that the equilibrium action a^* maximizing the dictator's utility is given by the following corner solutions:

$$a^* = \begin{cases} 0 & \text{if } \gamma_D \leq \frac{1}{3} \\ \rho\tau_D & \text{if } \gamma_D > \frac{1}{3}. \end{cases} \tag{4}$$

TP1 for $k = 1$ follows directly from (4).

Case $k > 1$. If $k > 1$, note that $U'(a) = -4 < 0$ on the interval $[\rho\tau_D, \infty)$. Therefore, the only candidate $\hat{a}_1$ for a maximizer of the dictator's utility function on the interval $[\rho\tau_D, \infty)$ is the corner solution $\rho\tau_D$, which is increasing in ρ and τ_D.

On the interval $[0, \rho\tau_D)$, note that the dictator's utility function is strictly concave as $U''(a) = -12^k (k-1) \gamma_D (\rho\tau_D - a)^{k-2} < 0$ for all $a \in [0, \rho\tau_D)$. First, assume that $U'(0) \leq 0$. Then, it follows from the concavity of the dictator's utility function that the $\hat{a}_2$ maximizing the dictator's utility function on the interval $[0, \rho\tau_D)$ is $\hat{a}_2 = 0$, which is independent of ρ.

Second, assume that $U''(0) > 0$, which happens for sufficiently high guilt aversion, i.e.,

$$\gamma_D > \frac{4}{12^k (\rho\tau_D)^{k-1}}.$$

Then, assuming a maximizer $\hat{a}_3$ exists on $[0, \rho\tau_D)$, it must be an interior solution given by the following first-order condition:

$$\hat{a}_3 = \rho\tau_D - \left(\frac{4}{\gamma_D 12^k}\right)^{\frac{1}{k-1}}.$$

It can be seen that $\hat{a}_3$ increases in ρ, in τ_D, and γ_D, which proves **TP1**.

Beliefs and Second-order Effects on Performance So far we have implicitly assumed that the dictator's second-order beliefs are constant at τ_D. However, as the dictator's equilibrium choice a^* weakly increases in ρ, first- and second-order beliefs must adjust accordingly. Hence, τ_R and τ_D and must be weakly increasing in ρ. This yields **TP2**.

As the dictator's equilibrium action is increasing in τ_D, this adjustment of beliefs leads to a (second-order) effect reinforcing **TP1**.

Opt-out Decision A (risk-neutral) recipient will opt in if

$$E[\pi_R] = 12E[\rho\tau_R(\rho)] > 2.$$

If $\mu = 0$, the dictator will choose $a = 0$ and the recipient's beliefs will adjust accordingly. Hence, $\tau_R(\cdot)$ and therefore also $E[\pi_R] = E[\rho\tau_R(\rho)]$ will be higher for $\mu = 1$ than for $\mu = 0$. As a result, opt-out rates are lower for $\mu = 1$ than for $\mu = 0$. This yields **TP3**.

Example: Assume that $\tau_R(\cdot) = 0$ for $\mu = 0$, and $\tau_R(\cdot) = 1$ for $\mu = 1$. Then we would have $0 > 2$ for $\mu = 0$ and $12\binom{1/2}{1/2}\binom{1/6}{5/6} = 6 > 2$ for $\mu = 1$.