

Broken Promises: An Experiment

Gary Charness and Martin Dufwenberg

August 6, 2008

Abstract: We test whether promises *per se* are effective in enhancing cooperative behavior in a form of trust game. In a new treatment, rather than permitting free-form messages, we instead allow only a bare promise-only message to be sent (or not). We find that bare promises are much less effective in achieving good social outcomes than free-form messages; in fact, bare promise-only messages lead to behavior that is much the same as when no messages are feasible. Our design also permits us to test the predictions of guilt aversion against the predictions of lying aversion. Our experimental results provide evidence that mainly supports the guilt-aversion predictions, but we also find some support for the presence of lying aversion.

Keywords: Behavioral economics, cheap talk, communication, cost-of-lying, credibility, guilt aversion, psychological game theory, promises

JEL codes: A13, B49, C72, C91, D63, D64, J41

Acknowledgements: We thank Andreas Blume, Vince Crawford, Tore Ellingsen, Håkan Holm, Navin Kartik, Vai-Lam Mui, Louis Putterman, Matthias Sutter, two referees, an associate editor, and participants in a seminar at the University of Central Florida and at the 2007 Arne Ryde Symposium on Communication in Games and Experiments at Lund University for very useful comments. We thank the National Science Foundation for support.

Contact: Gary Charness, University of California at Santa Barbara, charness@econ.ucsb.edu; Martin Dufwenberg, University of Arizona, martind@eller.arizona.edu.

1. INTRODUCTION

In the past couple of decades, there has been a tremendous upswing in interest in the effects of communication in economic situations. Theoretical work (e.g., Crawford & Sobel 1982; Green & Stokey 2007; Farrell 1987, 1988; Blume 1998) demonstrates that cheap talk may be effective when players' preferences are largely in alignment. Experimental work in coordination games, which feature such alignments, has found support (see, e.g., Cooper, Dejong, Forsythe, and Ross 1989, 1992; Charness 2000).

Yet payoffs are often not so well aligned. In this case there may be an incentive to make false statements, so that messages may lack credibility and communication is rendered ineffective. This is particularly true in environments such as Prisoner's Dilemma games, 'trust' games, and public-goods games, where the selfish action conflicts with the socially-optimal one. An important question is therefore what forms of communication are effective when a concern that others will act opportunistically makes cooperation difficult.

One positive note is that past studies have found that people have negative feelings about the use of deception (by oneself or by others), so that dishonesty is not always prevalent. In Brandts and Charness (2003), for example, people punish a selfish action much more frequently when it is preceded by a deceptive, self-serving message. And absent punishment opportunities, Gneezy (2005) finds that decision makers nevertheless curb the degree to which they engage in deception.

Promises are an integral part of deception. We examine whether a promise *per se* is effective in promoting cooperative behavior in our game. To do this, we alter the Charness and Dufwenberg (2006) (henceforth, "CD") design in a simple way: Rather than permitting free-form messages from the responder to the first mover, we only allow the responder to indicate whether

he makes a promise to play cooperatively. We find strong differences in behavior depending on whether messages are free-form or ‘pre-fabricated’ by the experimenter. In fact, behavior with pre-fabricated messages is similar to behavior (and the proportion of socially-optimal outcomes is much the same) when no communication is possible, in contrast with behavior with free-form promises.

There are two leading explanations concerning why people are reluctant to lie or to break promises. CD find support for a theory of *guilt aversion* in experiments involving a form of ‘trust’ game; this notion is based on so-called psychological game theory.¹ The stronger the responder believes that the first-mover believes that the responder will make the cooperative play, the more likely that the responder will indeed make this choice. This form of belief-dependent motivation provides a way for communication to shift behavior by moving beliefs so as to move motivation. CD provide evidence that promises (statements of intent) foster trust and cooperation by moving beliefs in a way that is consistent with guilt aversion.

An alternative and simpler explanation is that people experience a (belief-independent) cost of lying. We define a cost of lying as a discomfort (measurable in dollars) that depends on the truth-value of a claim, and that does not depend on other aspects of the claim than truth-value. In the context we consider (see section 2 for details), a player who makes any kind of promise to make a particular choice has to bear the cost of lying if he subsequently makes a different choice.

To the best of our knowledge, Ellingsen and Johannesson (2004) were the first to model this idea by introducing a “personal cost of being inconsistent,” and there are richer variants that still preserve belief-independence. Chen, Kartik & Sobel (2007) and Kartik (2008) develop

¹ See Geanakoplos, Pearce and Stacchetti (1989) and Battigalli and Dufwenberg (2007, 2008).

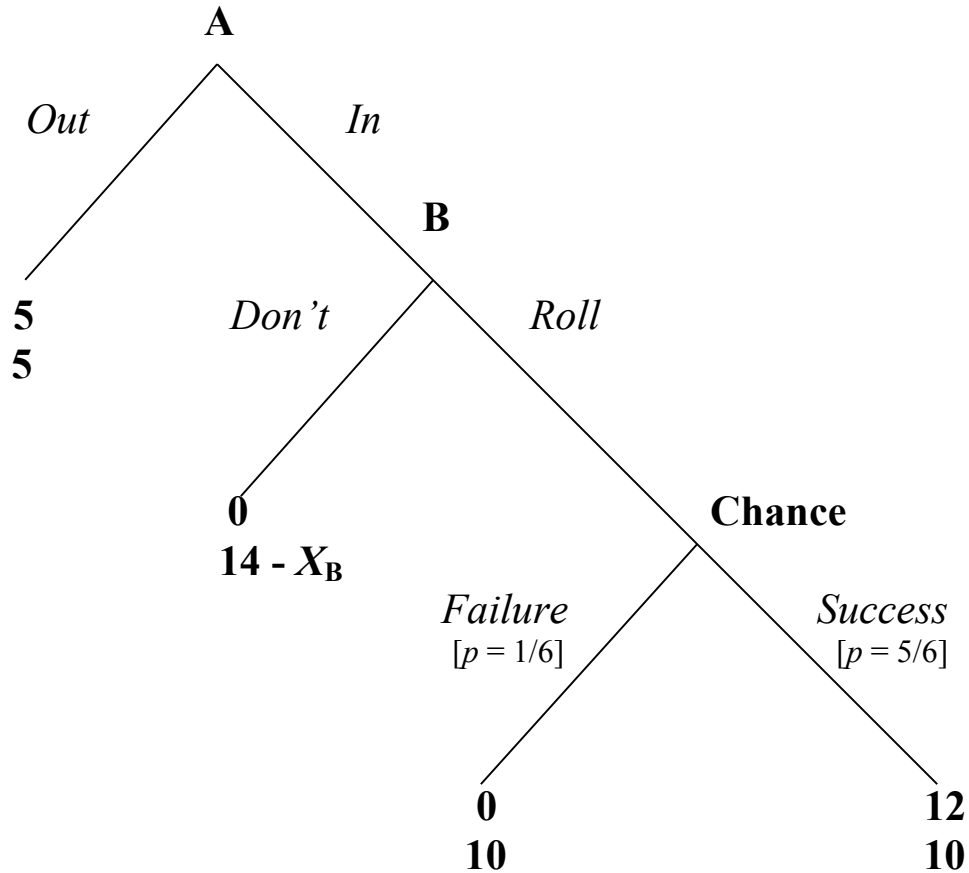
theory in which the cost of lying depends continuously on how far removed a statement is from the truth, and Gneezy (2005) presents evidence suggesting that an adequate account of costs of lying would need to make reference to how cost to vary with the payoff consequences of all parties involved.

The key distinction between these two alternative explanations concerns whether one needs to invoke belief-dependent motivation. Guilt aversion can provide a micro foundation for why people behave as they do, but it involves a more complex apparatus and to test for it requires somehow measuring beliefs. The simpler belief-independent cost-of-lying explanation (henceforth, “*lying aversion*”) may be preferred if it does a reasonable job of organizing the data. With guilt aversion, decision makers do not have preferences over the words they utter *per se*; promises, etc., matter only insofar as they move beliefs and thus motivation. On the other hand, lying aversion says that one experiences disutility from lying without regard to the effect that a lie has on beliefs. We develop hypotheses based on these theories, and test them with our data.

Section 2 presents the guilt-aversion and lying-aversion theories, and the hypotheses we test. The experimental design and results are described in section 3. Section 4 discusses some experimental evidence that is related by a focus on communication protocols in trust games rather than a link to guilt aversion versus belief-independent cost-of-lying. We conclude in section 5; in particular we discuss how one may reconcile our results with some recent work by Ellingsen, Johannesson, Tjøtta, and Torsvik (2007) and Vanberg (2008) that has called in question the empirical relevance of guilt aversion as a plausible means of explaining why promises foster trust and cooperation.

2. THEORY

In this section we summarize the two theories we focus on in this paper – guilt aversion and lying aversion – and state the hypotheses we propose to test. The following game tree depicts the strategic situation we focus on. The naming of players and strategies reflect the experimental design, to be described in detail in section 3 below.



We compare a no-communication treatment, which maps directly to the above game, and message treatments that augment the above game by adding a pre-play opportunity for B to send a message to A. CD reported on a no-communication treatment and on a treatment where B could send a free-form pre-play message to A; in addition we now consider allowing B to either send or not send a pre-fabricated bare promise-to-Roll to A.

The payoffs reflect dollar payments except for X_B , which is a psychic cost to player B if the profile $(In, Don't)$ is played. We take X_B to be measurable in dollars too, so $14-X_B$ is a dollar number. The two theories we contrast and test differ in regards to how they determine X_B :

Guilt Aversion

This is the sentiment in focus in CD. If B chooses *Don't* he feels guilt X_B in proportion to how many dollars he believes he is costing A relative to what A expects to get. Formally, suppose A chooses *In* and let $\pi_A \in [0,1]$ be the probability that A then assigns to B choosing *Roll*. When B moves, he has a belief regarding π_A represented as a probability measure; let $\pi_B \in [0,1]$ denote its mean. If A chooses *In* she thus believes she will get $\pi_A \cdot [(5/6) \cdot 12 + (1/6) \cdot 0] = 10 \cdot \pi_A$, and B believes A believes she will get $10 \cdot \pi_B$. If B chooses *Don't*, A gets 0 and B experiences guilt in proportion to the difference: $10 \cdot \pi_B - 0 = 10 \cdot \pi_B$. Letting $\gamma_B \geq 0$ be a constant measuring B's sensitivity to guilt, in the above game tree we get $X_B = \gamma_B \cdot 10 \cdot \pi_B$. If $\gamma_B \cdot 10 \cdot \pi_B > 4$, a rational B will *Roll*.

Guilt aversion provides a route by which promises may foster trust and cooperation. By making a promise to *Roll*, B may boost π_A and A's subjectively expected payoff to choosing *In*. This is incentive-compatible and plausible, because if B believes a promise will boost π_A then a promise will boost π_B and strengthen the incentives for B to *Roll* (as X_B goes up), which justifies the boost to π_A . A promise may thus feed a self-fulfilling circle of beliefs and beliefs about beliefs that the parties will play $(In, Roll)$ rather than $(Out, Don't)$.

However, the guilt aversion hypothesis is flexible enough that this effect may or may not occur. Guilt aversion allows, for example, that while the effect occurs with a 'full-blooded' and

carefully worded promise, it will not occur with a pre-fabricated bare promise-to-*Roll*. That a pre-fabricated bare promise-to-*Roll* does *not* boost β may be plausible; if B believes a pre-fabricated bare promise-to-*Roll* will not boost β then it will not boost β and so not strengthen the incentives for B to *Roll*, which justifies no boost to β . Guilt aversion thus can accommodate that only a full-blooded promise feeds a self-fulfilling circle of beliefs and beliefs about beliefs that (*In, Roll*) rather than (*Out, Don't Roll*) will be played. Our design allows us to measure beliefs so that we can explore this possibility.

Lying Aversion

Belief-independent lying aversion applied to the above game means that X_B is a scalar.² If pre-play communication is not possible, there can be no lie and $X_B = 0$. The predicted outcome is the profile (*Out, Don't*). If, on the other hand, pre-play communication is possible then $X_B > 0$ will apply whenever B issues a promise to *Roll*. If $X_B > 4$, B would subsequently *Roll*, and if A figures this to be likely she would best respond by choosing *In*. So B will want to issue a promise because otherwise play would be (*Out, Don't*), and he would get \$5 rather than \$10. Lying aversion would thus foster trust and cooperation when communication is possible.³

The argument of the previous paragraph makes no distinction with respect to the nature of a lie. A lie is a lie is a lie.... Our tests for lying aversion relates to such a perhaps blunt view

² In richer settings we would need to be more nuanced. In games with more than two strategies, the cost of lying may depend continuously on how far removed from the truth a statement is (as modeled by Chen, Kartik and Sobel 2007 and by Kartik 2008). If lying costs were compared across games, then those costs might reasonably vary with the payoff consequences for all players involved (as evidenced by Gneezy 2005). Both these points are moot in our binary-choice game form with given monetary consequences.

³ Since distributional social preferences may also affect choices it is not necessary for X_B to be greater than 4 to affect behavior. For example, consider a B whose social preferences would lead him to choose *Roll* (without communication) if her material payoff from doing so were 11, but not if his payoff from *Roll* is 10. Then his choice would change if $X_B > 1$.

on the essence of lying aversion; the cost of a lie is the same whether the corresponding promise is full-blooded & carefully worded or pre-fabricated and bare.⁴

Hypotheses

According to standard theory, player B should always choose *Don't Roll*; knowing this, A should always choose *Out*. This holds whether or not communication is feasible. Of course, some models of social preferences (e.g., Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Charness and Rabin 2002) allow for the possibility that B could prefer the expected outcome of (10,10) to the outcome (0,14) due to either difference aversion or quasi-maximin preferences. Nevertheless, none of these models predict any effect from communication in this setting. This leads us to the two candidate explanations of guilt aversion and lying aversion. Below we list several predictions of these competing stories regarding the relationship of beliefs and choices, with free-form promises and with bare promises.

First (**H1**), lying aversion predicts that the likelihood of B's choosing *Roll* will be uncorrelated with B's beliefs concerning A's beliefs; this should hold in both communication treatments.⁵ Alternatively (**H1a**), the bedrock prediction for guilt aversion is that there will be a significant positive correlation between B's second-order beliefs and the likelihood that B's choose *Roll*.

⁴ One could conceive of a more refined notion of lying aversion such that, for example, a cost of lying is experienced following full-blooded promises but not bare ones. Such a pattern of sentiments should presumably be called something like "aversion to full-blooded lying" rather than lying aversion.

⁵ This holds if players' costs of lying parameters are uncorrelated with their second-order beliefs. Our conclusions are conditional on that assumption, and we note that it is not completely innocuous. A kind of 'false consensus' effect, whereby B players cue their second-order beliefs on the own choice, could nullify it. See CD section 5.3 for related discussion, and Ellingsen *et al* (2007) and Al-Ubaydli and Lee (2008) for evidence indicating that this concern may have empirical merit.

Second (**H2**), lying aversion predicts that A's will be equally likely to choose *In* conditional upon receiving either a free-form or impersonal promise. On the other hand (**H2a**), guilt aversion does not take an *a priori* stand on which messages have cutting power, and we focus on the seemingly natural possibility that A's will be more likely to choose *In* after receiving a free-form promise. If impersonal promises are completely ineffective, we would expect that the rate of *In* is independent of whether or not such a message was sent and this could still be consistent with guilt aversion (as long as **H1a** is also supported).

Third (**H3**), lying aversion predicts that B's will have the same probability of keeping a promise regardless of whether this was a free-form or pre-fabricated promise. The alternative possible prediction (**H3a**) for guilt aversion that we emphasize is that if B's are guilt averse and anticipate that bare-bones promises affect beliefs less than do personalized ones,⁶ the probability that B's keep their promises will be higher with free-form messages.

Fourth (**H4**), lying aversion predicts that A's should have the same beliefs, conditional upon receiving a promise, in both communication treatments. The guilt-aversion alternative hypothesis (**H4a**) is that A's will hold higher beliefs about the likelihood of B's choosing *Roll* when receiving a free-form promise.⁷

3. EXPERIMENT

Design

As with CD, sessions were conducted at UCSB, in a large classroom divided into two sides by a center aisle. Participants were seated at spaced intervals. In addition to the sessions

⁶ As with **H2a**, there is nothing *per se* in guilt aversion that would lead B to anticipate this difference. Thus, even with guilt aversion present, we might not observe that B's are more likely to keep their promises with free-form messages. **H3a** is the natural complement to **H2a**.

⁷ Again, this is not implied by, but is consistent with, guilt aversion and is a natural complement to **H2a** and **H3a**.

conducted in that study, we added three sessions with 26-36 participants per session; there were 96 participants in these three sessions. No one could participate in more than one session. Average earnings were about \$14 (including a \$5 show-up fee); sessions took about one hour.

In each session, participants were referred to as “A” or “B”. A coin was tossed to determine which side of the room was A and which was B. Identification numbers were shuffled and passed out face down, and participants were informed that these numbers would be used to determine pairings (one A with one B) and to track decisions for payoffs.⁸

The outcomes and corresponding payoffs were described to the participants in this chart:

	A receives	B receives
A chooses <i>Out</i>	\$5	\$5
A chooses <i>In</i> , B chooses <i>Don't Roll</i>	\$0	\$14
A chooses <i>In</i> , B chooses <i>Roll</i> , die = 1	\$0	\$10
A chooses <i>In</i> , B chooses <i>Roll</i> , die = 2,3,4,5, or 6	\$12	\$10

In the absence of communication, A first chooses *In* or *Out*; next, B chose whether to *Roll* or *Don't Roll* a 6-sided die. B made this choice without knowing A's actual choice, but the instructions explained that B's choice would be immaterial if A chose *Out*. As in CD, we thus obtain an observation for every B (“the strategy method”). The outcome corresponding to a successful project occurred only if the die came up 2, 3, 4, 5, or 6 after a *Roll* choice.⁹ After the decisions had been collected, a 6-sided die was rolled for each B; this was made clear to the participants in advance, to avoid the anticipated loss of public anonymity for B's who chose *Don't Roll*. This roll was determinative if and only if (*In*, *Roll*) had been chosen.

⁸ The complete instructions for the bare-promise-only treatment are presented in Appendix A.

⁹ Notice that principals therefore face a choice between a sure thing and a lottery. To the extent that risk preferences come into play, this could affect a principal's choice. However, our interest is in the difference in behavior across treatments, rather than the levels *per se*; if risk preferences are the same across independent draws from the subject population, they should not affect this difference.

In our new communication treatment, B transmitted a message to A before the choice of *In* or *Out* (as in CD). Instead of free-form messages, each agent was given two additional sheets of paper. One of these stated: “I promise to choose *Roll*,” while the other was blank; a promise was not binding. All of this was public information. The agent placed one of these two sheets in the envelope provided for this purpose. We conveyed this envelope to the appropriate A, and then A and B proceeded as described above.¹⁰

After we collected the strategic choices, we passed out decision sheets that invited participants to make guesses about their counterparts, and offered to reward good guesses. A’s were asked to guess the proportion of B’s who chose *Roll*.¹¹ Knowing that A’s made this guess, B’s were then asked to guess the average guess made by A’s who chose *In*. These guesses (by A’s and B’s) constitute the data we take to represent the players’ beliefs (the probability that A assigns to *Roll* and the belief of B (conditional on *In*) about the probability that A assigns to *Roll*). If a guess was within five percentage points of the realization, we rewarded the guesser with \$5 (we also told participants that we would pay \$5 for all B guesses if no A’s had chosen *In*).¹² As our game is one-shot and we didn’t mention guesses until after strategies were chosen, the belief elicitation should not affect participants’ prior choices.

¹⁰ In the bare-promise treatment, we added a line stating that the promise was not binding. We felt that this was needed to avoid a high rate of promise-keeping due to misunderstanding. While there was no such explicit statement in the original instructions (stating that promises were not binding might have suggested to B’s that they were supposed to be making promises), we typically made it clear to the group (in verbal responses to questions) that B’s could make either choice regardless of what they had written. Bochet and Putterman (2007) provide a thorough discussion of the issues involved with the design problem of having to explicitly indicate that a promise isn’t binding.

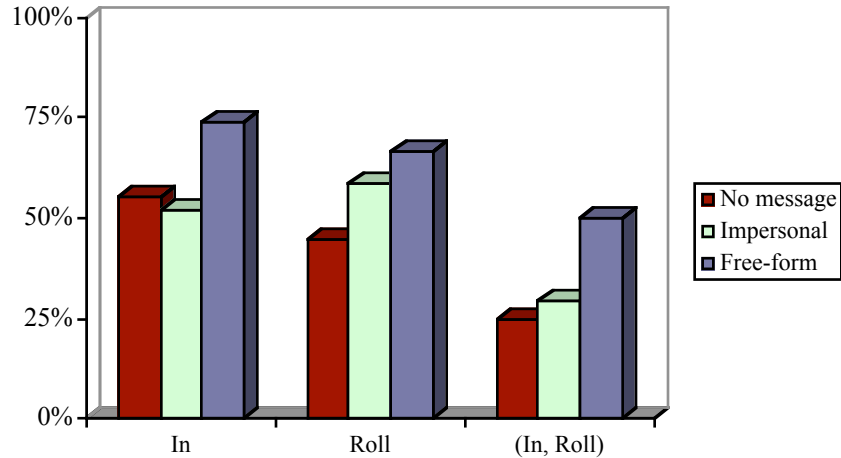
¹¹ We did not ask A’s to guess the probability the paired B would choose *Roll*, as we don’t observe this likelihood. The observed binary choice would make this simply a *Yes* or *No* guess.

¹² How to best elicit beliefs is itself an important issue. Our scheme has the virtue of being simple to describe in instructions (as well as of staying true to CD). We refer to Andersen, Fountain, Harrison & Rutström (2007) for an in-depth discussion about the pros and cons of various methods.

Results

We observe substantially more cooperative behavior in the free-form message treatment than in the promise-only and no-promise treatments. Figure 1 summarizes:

Figure 1 - Message Types and Behavior



The rates of cooperative play in the impersonal-promise treatment are generally much closer to the rates without messages than to the rates with free-form messages; an exception is the B impersonal-promise *Roll* rate, which is intermediate. Table 1 confirms these visual results:

Table 1: Results Across Treatments

	No Communication	Impersonal Promises	Free-form messages
% In	25/45 (55.6%)	25/48 (52.1%)	31/42 (73.8%)
% Roll	20/45 (44.4%)	28/48 (58.3%)	28/42 (66.7%)
% (In, Roll)	24.7%	29.2%	50.0%

Overall, the observed likelihood of the (*In, Roll*) outcome was 29.2% when only impersonal promises are feasible with impersonal promises and 50.0% when free-form messages

are feasible. This compares to the baseline of 24.7% without communication.¹³ Thus, impersonal communication only increases the likelihood of the (*In*, *Roll*) choice by 4.5 percentage points over no communication, while free-form communication increases this rate by 25.3 percentage points.¹⁴ The test of the difference in proportions (Glasnapp & Poggio 1985) indicates that there is no significant difference between the rates of (*In*, *Roll*) outcomes in the no-communication and impersonal-promises treatments ($Z = 0.49$), but that the rate with free-form communication is significantly higher than with either no communication ($Z = 2.44$, $p = 0.015$, two-tailed test) or impersonal communication ($Z = 2.02$, $p = 0.044$, two-tailed test).

The same test shows that A's choose *In* is significantly less often with impersonal promises than with free-form messages, ($Z = 2.12$, $p = 0.034$, two-tailed test); however, while B's are more likely to choose *Roll* with free-form messages, the difference is not significant ($Z = 0.81$). Comparing behavior with impersonal promises to behavior with no communication, there is no significant difference for *In* rates ($Z = -0.34$). Finally, B's in the impersonal-promises treatment have slightly greater *Roll* rates than B's in the no-communication treatment; however the difference is not statistically significant ($Z = 1.34$, $p = 0.180$, two-tailed test; if we consider only B's who sent impersonal promises, this difference is a bit larger ($Z = 1.53$, $p = 0.126$).¹⁵

We next consider the comparative effects of participant-generated and experimenter-generated promises, which bear more directly on our hypotheses. Table 2 shows A and B

¹³ This 24.7% is the predicted rate, rather than the actual observed rate, as the actual combinations are entirely random without communication. We use the actual rate when there was communication.

¹⁴ As seen in Table 2 below, the rate of (*In*, *Roll*) outcomes grows to 66.7% (a 42.0 percentage-point increase) when free-form promises are made.

¹⁵ To the degree that this is independent of B's beliefs (and the average guess of the seven B's who did not send a bare promise was very close to that of the other 41 B's who did) according to whether they sent promises, this suggests some support for lying aversion.

behavior after a promise in each treatment,¹⁶ and Table 3 shows average beliefs for A's who either did or didn't receive a promise and for all B players, according to whether they chose *Roll*:

Table 2: Behavior after Promises

	Impersonal promises	Free-form promises
% In	23/41 (56.1%)	22/24 (91.7%)
% Roll	25/41 (61.0%)	18/24 (75.0%)
% (In, Roll)	13/41 (31.7%)	16/24 (66.7%)

Table 3: Average Beliefs, by Category and Treatment

	Impersonal	Free-form
B's choosing <i>Roll</i>	59.3 [3.3]	73.2 [3.0]
B's choosing <i>Don't Roll</i>	50.6 [4.6]	45.1 [7.2]
A's receiving promises	49.7 [5.1]	65.8 [5.1]
A's not receiving promises	48.1 [12.7]	50.0 [7.2]

Standard errors are in brackets.

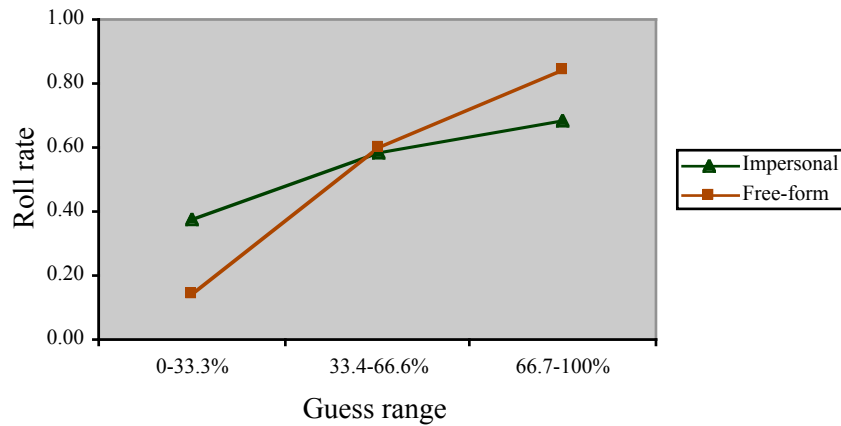
Let us consider the predictions of guilt aversion and lying aversion in the light of these data. First, does the likelihood of *Roll* depend on B's second-order beliefs? There is a strong relationship with free-form messages, as the Wilcoxon rank sum test gives $Z = 3.20$ ($p = 0.001$, one-tailed test¹⁷) and a simple OLS regression with B's second-order beliefs as the dependent variable and a dummy for *Roll* as the explanatory variable gives $t = 4.30$ ($p = 0.000$) for the coefficient on *Roll*. This relationship is weaker in the bare-promise treatment; the Wilcoxon test gives $Z = 1.46$ ($p = 0.072$) and the corresponding regression gives $t = 1.59$ ($p = 0.059$) for the

¹⁶ We note that the rate of promises is 41/48 (85%) in the bare-promise treatment compared to 24/42 (57%) in the free-form case; these rates are significantly different ($Z = 2.99$, $p < 0.01$). Most likely this difference is due to the fact that B's in the free-form treatment had to come up with the idea of making a promise. While it might seem that it is less costly to simply turn in a bare promise than to write a message, we point out that a higher percentage of B's (38/42, or 90%; 10 of these messages were classified as "empty talk") in the free-form treatment did send written messages. There is no significant difference between the rates of message-sending in the two treatments ($Z = -0.73$).

¹⁷ We use one-tailed tests for our directional hypotheses.

Roll coefficient. Thus, with respect to the first hypothesis, we find some support for the guilt-aversion predictions over those of lying aversion. Figure 2 visually displays the relationship between B's guesses and *Roll* rates:

Figure 2: Likelihood of Roll, by B Guess



We next test whether A's are equally likely to choose *In* after receiving impersonal and free-form promises. Table 2 shows that A's are much more likely to choose *In* after free-form promises, and this is confirmed by the test of proportions ($Z = 3.00, p = 0.001$). In addition, there is no difference between *In* rates in the no-communication treatment and with impersonal promises ($Z = 0.05$). Thus, this test provides clear evidence in favor of the guilt-aversion alternative hypothesis H2a.

The third hypothesis relates to whether B is equally likely to keep a promise in the two treatments. Table 2 shows that the rate of promise-keeping is slightly higher with free-form promises (75% versus 61%); however, this difference is not statistically significant (the test of proportions gives $Z = 1.15, p = 0.125$).

Finally, guilt aversion predicts that A's who receive free-form promises have higher beliefs about the likelihood that B's choose *Roll* than do A's who receive impersonal promises;

lying aversion predicts no such relationship. Table 3 shows that A's beliefs are remarkably similar for A's who do not receive a free-form promise, but that average A's beliefs are substantially higher when they receive such a promise. The Wilcoxon test gives $Z = 1.69$ ($p = 0.046$) and a simple OLS regression with A's beliefs after receiving a promise as the dependent variable and a dummy for the free-form treatment as the explanatory variable gives $t = 1.92$ ($p = 0.030$) for the coefficient on the treatment. Related to this issue, we can examine the distribution of beliefs for A's who receive either impersonal or free-form promises. If these distributions were similar, then this would be support for lying aversion. However, a Kolmogorov-Smirnov test (using five equal ranges) gives $\chi^2 = 5.08$, $p = 0.024$, indicating that these distributions are significantly different.

How can we summarize the evidence? Of the five (two for H1) predictions for lying aversion, three are rejected with statistical significance in favor of the guilt-aversion alternative hypothesis, one is only rejected at marginal statistical significance, and in one case (H3) we cannot reject the lying-aversion hypothesis. While this is something of a mixed bag, we note that all five of the differences were in the direction predicted by guilt aversion, which would happen randomly 3.1% of the time. On balance, we interpret the evidence as providing more support for guilt aversion than for lying aversion.

4. PREVIOUS WORK ON COMMUNICATION PROTOCOLS IN TRUST GAMES

Our experimental design is developed with the goal of empirically comparing two theories for why communication may foster trust and cooperation: guilt aversion versus belief-independent cost-of-lying. As a by-product, however, we also end up producing results that relate to the literature that open-mindedly investigates the impact of various forms of communication protocols

in various forms of trust games, without necessarily testing any preconceived hypotheses. In this section we briefly discuss this literature. Note that the studies we discuss have not measured second-order beliefs, so they do not speak directly to the guilt-aversion theory that we address.

Glaeser, Laibson, Scheinkman & Soutter (2000; see pp. 821, 830) make an across-treatment comparison of the effect of an experimenter-specified promise opportunity. In one condition, they require the recipient in a trust game to check either a box stating that they promise to return at least as much as will be sent or a box stating that no promise is being made. Their promise condition seems to anchor responses, as more responders return exactly as much as they were sent in the promise condition than in the no-promise condition. However, in contrast to CD, they do not see an overall boosting effect by promises on trust and trustworthiness; both more generous and more selfish choices were less likely, and the average return ratio did not vary across these treatments.

Andreoni (2005) also tests behavior in a form of a trust game. In the “non-binding” condition, the trustee can elect to give the first-mover the option of restoring payoffs to the original endowment. If so, and if the trustor exercises this option after observing the trustee’s choice, the trustee can choose (by checking a box) whether to in fact restore the payoffs to the original endowment (“satisfaction guaranteed”). But this implicit promise is only partially effective. While responders indeed make more favorable choices when they have made promises, first-movers don’t trust the claim.¹⁸

Ben-Ner, Putterman & Ren (2007) find results that bear similarity to ours. Participants interact in trust games with two different forms of pre-play communication: numerical (tabular) only, or both verbal (in a chat box) and numerical. While either form increases trusting and/or

¹⁸ This is similar to the results in our bare-promise treatment, as in both cases the behavior of the promise-maker is more affected by the promise than is the behavior of the receiver of the promise.

trustworthiness, when verbal communication is included there is a substantially larger effect. Both trusting and trustworthiness are enhanced when the parties can use words, particularly when an agreement is reached with words and not only with the exchange of numerical proposals. In a follow-up study by Bochet and Putterman (2007), some treatments included an option to send a pre-specified promise to contribute an amount in a VCM (participants filled in the amount). In this case, bare promises were considerably less effective than free-form ones, except in treatments where there is also an option for punishing broken promises.

These studies thus share with ours the insight that the effect of bare promises on trust and cooperation tends to be meager. We hasten to add that one must not jump to the conclusion that bare, impersonal messages are always ineffective in games; as mentioned earlier, such messages have been shown to produce strong results in coordination games where the players' interests are aligned. But it is considerably more difficult to induce players to play strategies that are dominated in terms of own payoffs. To move beliefs in games where payoffs are not aligned, such as the trust and Prisoners' Dilemma games mentioned earlier, a more personalized form of communication seems to be necessary, or is at least much more effective.

Apart from the evidence we report in this paper, some relevant further evidence for this view comes from Charness (2000) where there is a dramatic difference between the effectiveness of one-way impersonal messages in the Stag Hunt and in the Prisoners' Dilemma.¹⁹ Thus to enhance cooperation with pre-play communication when there is a dominant strategy in own payoffs, it appears that there must be something in the message beyond the words used *per se*. A player must sound like he really means what he says; if not, another player's beliefs won't be favorably affected, and his words won't foster trust and cooperation.

¹⁹ Such messages were extremely effective in the Stag Hunt, as the likelihood of a Stag choice after this message was 94% (this compares to 35% without messages). However, in the Prisoners' Dilemma, the probability of the cooperative choice after this message was sent was only 10%.

5. CONCLUSION

We find that impersonal promises that are pre-fabricated by the experimenter are much less effective than free-form messages generated by players in a form of trust game. Even though such bare and impersonal messages have been shown to quite effective in coordination games, it seems that more is needed to achieve cooperative behavior when selfish behavior leads to poor social outcomes. In this case, an endogenous promise seems to be much more effective.

That people sometimes do not like to lie is clear from introspection and is supported by experimental data from e.g. from Gneezy (2005), Sutter (2007), and Hurkens and Kartik (2007). How can this be explained? CD suggest that this is the result of guilt aversion, the view that it is not the words *per se* about which decision makers care, but rather the degree of expectation mismatch that these words instill. More precisely, a person who lies and anticipates that he successfully influences others' beliefs suffers from guilt to the extent that he does not live up to these elevated beliefs. This provides a disincentive to lie and a complementary objective to issue promises in order to gain commitment power in contexts where these statements would be believed.²⁰

An alternative view is that there is a belief-independent cost-of-lying that is either constant for each individual or depends on the payoff consequences for the people involved.²¹ In fact, some recent studies cast some doubt on the guilt-aversion explanation. Ellingsen *et al.* (2007) find no significant relationship between beliefs and action in a dictator game, a trust

²⁰ It is crucial to note the stated caveats here: Lies matter to motivation only if they “successfully influence others’ beliefs” and in “contexts where these statements would be believed”.

²¹ As previously noted in footnote 2, since we hold the payoff consequences constant in our treatments, these two formulations are equivalent with respect to our study. In addition, we point out that guilt aversion actually handles a cost of lying that depends on payoff consequences quite well: A guilt-averse decision maker feels guilt if he does not live up to the expectations of someone else. The degree of guilt he feels is proportional to the dollar difference between (his beliefs about) what the other player believes he will get and what he actually gets. This is built into the specification of CD, and in the follow-up paper by Battigalli and Dufwenberg (2007) that extends guilt aversion to apply not only to trust games but also to a large class of extensive games.

game, and a game similar to our own. Based on their evidence, they conclude that guilt aversion is not present. They interpret the correlation between B's beliefs and B's behavior in CD as instead reflecting a "false consensus effect".²²

Vanberg (2007) conducts an experiment using a modified dictator game with bilateral communication and subsequent random determination of who divides the money. He adds the possibility that the dummy player will be matched with a different dictator than the person who received the message. Overall, there is a strong positive relationship between beliefs and action (significant in all regressions in Table A1 of Appendix 3), in principle supporting guilt aversion. However, there is also a strong effect from participant-generated promises even when this is taken into account, and guilt aversion cannot explain his result (Tables 2 and 3) that B's beliefs increase when B has sent a promise but is paired with a different A (who reads the message that B had sent), but *Roll* rates don't increase in this case. Vanberg interprets his evidence as indicating that actions can induce guilt for reasons other than expectations or beliefs.²³

The main purpose of CD was to test for guilt aversion, but in section 5.2 we argued against a fixed cost of lying being a major factor in our data. In light of the evidence provided by Ellingsen *et al.* and Vanberg, a more balanced view would suggest that guilt aversion describes one important aspect of human motivation, and that lying aversion describes another, with neither explanation providing the whole picture. Note that to the extent that lying aversion or something similar is present, it has somewhat limited applicability. For example, lying aversion is relevant only in contexts where communication is possible, whereas guilt aversion applies more generally and many, many experiments have demonstrated that non-selfish choices

²² There are some design differences that could possibly affect behavior, particularly in their replication of our game; see Al-Ubaydli and Lee (2008) for a discussion.

²³ He states: "Future attempts to theoretically account for the influence of guilt must acknowledge the independent effect that contracts and other sources of obligation can have on human motivation."

happen even when communication is ruled out. Moreover, there are contexts where lying seems more-or-less expected and where liars seem not to suffer much.²⁴

The data from our bare-promises treatment refines the story further. Promises *per se* are not particularly effective in promoting cooperative behavior in a form of the trust game. Personalized promises are much more effective than content-free promises, even when the underlying payoff structure of the game is identical. We examine the distinct predictions following from guilt aversion and lying aversion and in most comparisons we can reject the predictions for lying aversion at statistically-significant levels. In all cases, behavior is directionally consistent with guilt aversion, which allows that only particular kinds of (full-blooded!) promises move beliefs and foster trust and cooperation. If you sound like you mean it, the chances are greater that you will do it.

It is clear that the jury is still out regarding why people make or break promises, that more work needs to be done, and that a good answer may well be multi-faceted. Opening up the black box of communication seems one of the more exciting directions for behavioral economics, and we hope that our results offer something for other researchers to build upon.

²⁴ CD cite chit-chat around the poker table, and give a poker textbook reference. Other environments include used car sales, promises made by politicians, tax returns sent to the IRS, and testimony in traffic courts (under oath!).

REFERENCES

- Al-Ubaydli, O. and M. Lee (2008), "Guilt aversion vs. intentional reciprocity in a trust game with punishment," mimeo.
- Andersen, S., J. Fountain, G. Harrison & E. Rutström (2007), "Eliciting Beliefs: Theory and Experiments," Working Paper 07-08, Department of Economics, College of Business Administration, University of Central Florida.
- Andreoni, J. (2005), "Trust, Reciprocity, and Contract Enforcement: Experiments on Satisfaction Guaranteed," mimeo.
- Battigalli, P. and M. Dufwenberg (2008), "Dynamic Psychological Games," forthcoming, *Journal of Economic Theory*.
- Battigalli, P. and M. Dufwenberg (2007), "Guilt in Games," *American Economic Review, Papers & Proceedings*, **97**, 170-76.
- Ben-Ner, A., L. Putterman & T. Ren (2007), "Lavish Returns on Cheap Talk: Non-binding Communication in a Trust Experiment," mimeo.
- Bochet, O. and L. Putterman (2007), "Not Just Babble," mimeo.
- Bolton, G. and Ockenfels, A. (1999), "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, **90**, 166-193.
- Brandts, J. and G. Charness (2003), "Truth or Consequences: An Experiment," *Management Science*, **49**, 116-130.
- Charness, G. (2000), "Self-serving Cheap Talk and Credibility: A Test of Aumann's Conjecture," *Games and Economic Behavior*, **33**, 177-194.
- Charness, G. and M. Rabin (2002), "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, **117**, 817-869.
- Charness, G. and M. Dufwenberg (2006), "Promises and Partnership," *Econometrica*, **74**, 1579-1601.
- Chen, Y., N. Kartik & J. Sobel (2007), "Selecting Cheap-Talk Equilibria," forthcoming in *Econometrica*.
- Cooper, R., D. DeJong, R. Forsythe & T. Ross (1989), "Communication in the Battle of the Sexes Game: Some Experimental Results," *RAND Journal of Economics*, **20**, 568-587.
- Cooper, R., D. DeJong, R. Forsythe & T. Ross (1992), "Communication in Coordination Games," *Quarterly Journal of Economics*, **107**, 739-771.
- Crawford, V. (1998), "A Survey of Experiments on Communication via Cheap Talk," *Journal of Economic Theory*, **78**, 286-298.
- Crawford, V. and J. Sobel (1982), "Strategic Information Transmission," *Econometrica*, **50**, 1431-1452.
- Ellingsen, T. & M. Johannesson (2004) "Promises, Threats, and Fairness", *Economic Journal* **114**, 397-420.
- Ellingsen, T., M. Johannesson, S. Tjøtta & G. Torsvik (2007), "Testing Guilt Aversion," mimeo.

- Fehr, E. and K. Schmidt (1999), "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, **114**, 817-868.
- Geanakoplos, John, David Pearce & Ennio Stacchetti (1989), "Psychological Games and Sequential Rationality", *Games & Economic Behavior*, **1**, 60–79.
- Glaeser, E., D. Laibson, J. Scheinkman & C. Soutter (2000), "Measuring Trust," *Quarterly Journal of Economics*, **115**, 811-846.
- Glasnapp, D. and J. Poggio (1985), *Essentials of Statistical Analysis for the Behavioral Sciences*, Merrill: Columbus.
- Gneezy, U. (2005), "Deception: The Role of Consequences," *American Economic Review*, **95**, 384-394.
- Green, J. and N. Stokey (2007), "A Two-Person Game of Information Transmission," *Journal of Economic Theory*, **135**, 90-104.
- Hurkens, S. and N. Kartik (2007), "Would I lie to you? On social preferences and lying aversion," mimeo.
- Kartik, N. (2008), "Strategic Communication with Lying Costs," mimeo.
- Sutter, M. (2007), "Deception through telling the truth?! Experimental evidence from individuals and teams," forthcoming in *Economic Journal*.
- Vanberg, C. (2007), "Why Do People Keep Their Promises: An Experimental Test of Two Explanations," forthcoming in *Econometrica*.

APPENDIX A – INSTRUCTIONS

Thank you for participating in this session. The purpose of this experiment is to study how people make decisions in a particular situation. Feel free to ask us questions as they arise, by raising your hand. Please do not speak to other participants during the experiment.

You will receive \$5 for participating in this session. You may also receive additional money, depending on the decisions made (as described below). Upon completion of the session, this additional amount will be paid to you individually and privately.

During the session, you will be paired with another person. However, no participant will ever know the identity of the person with whom he or she is paired.

Decision tasks

In each pair, one person will have the role of A, and the other will have the role of B. The amount of money you earn depends on the decisions made in your pair.

On the designated decision sheet, each person A and person B will indicate whether he or she wishes to choose IN or OUT. If A chooses OUT, A and B each receives \$5. We will collect these sheets after the choices have been indicated. Next, each person B will indicate whether he or she wishes to choose ROLL or DON'T ROLL (a die). Note that B will not know whether A has chosen IN or OUT; however, since B's decision will only make a difference when A has chosen IN, we ask B's to presume (for the purpose of making this decision) that A has chosen IN.

If A has chosen IN and B chooses DON'T ROLL, then B receives \$14 and A receives \$0. If A has chosen IN and B chooses ROLL, then B receives \$10 and rolls a six-sided die to determine A's payoff. If the die comes up 1, A receives \$0; if the die comes up 2-6, A receives \$12. (All of these amounts are in addition to the \$5 show-up fee.) This information is summarized in the chart below:

	A receives	B receives
Either A or B chooses OUT	\$5	\$5
A and B choose IN, B chooses DON'T ROLL	\$0	\$14
A and B choose IN, B chooses ROLL, die=1	\$0	\$10
A and B choose IN, B chooses ROLL, die=2,3,4,5, or 6	\$12	\$10

A Promise

Prior to the decision by A and B concerning IN or OUT, B has an option to promise A that he or she will choose ROLL if A chooses IN. Each B has been given two additional sheets of paper. One sheet has the statement: "I promise to choose ROLL." If you wish to make a promise, please circle this statement. The other is blank, except for the letter B on top. Please return one of these sheets face down to the experimenter, who will convey it to the appropriate A participant, and then A and B will proceed as described above. B may still choose to ROLL or DON'T ROLL after a promise.