Would I Lie To You? On Social Preferences and Lying Aversion[*]

Sjaak Hurkens[†] Navin Kartik[‡]

First version: May 2006; Current version: May 26, 2008

Abstract

This paper reinterprets the evidence on lying or deception presented in Gneezy (2005, *American Economic Review*). We show that Gneezy's data are consistent with the simple hypothesis that people are one of two kinds: either a person will never lie, or a person will lie whenever she prefers the outcome obtained by lying over the outcome obtained by telling the truth. This implies that so long as lying induces a preferred outcome over truth-telling, a person's decision of whether to lie may be completely insensitive to other changes in the induced outcomes, such as exactly how much she monetarily gains relative to how much she hurts an anonymous partner. We run new but similar experiments to those of Gneezy in order to test this hypothesis. While we also confirm that there is an aversion to lying in our subject population, our data cannot reject the simple hypothesis described above either.

Keywords: experimental economics, lying, deception, social preferences

J.E.L. Classification: C91

[*] We thank Vince Crawford for encouragement and advice; Nageeb Ali, Jordi Brandts, Gary Charness, Arnaud Costinot, Nir Jaimovich, Ivana Komunjer, Albert Satorra, Roberto Weber, Tim Cason (the Editor) and two anonymous referees for comments and discussions; and various seminar and conference audiences for feedback. We also thank Uri Gneezy for generously clarifying unpublished details about his experiments, and for bringing the work of Matthias Sutter to our attention. This research was initiated when Sjaak Hurkens visited the Department of Economics at UCSD, and revisions were undertaken when Navin Kartik was at the Institute for Advanced Study at Princeton; we are grateful to these institutions for their hospitality. Financial support from grant PR2005-0204 by the State Secretary of Universities and Research of the Spanish Ministry of Education and Science (MEC), grant CONSOLIDER-INGENIO 2010 (CSD2006-00016), the Institute for Applied Economics at UCSD, the Institute for Humane Studies, and the Russell Sage Foundation is gratefully acknowledged. Our apologies to Charles and Eddie and Eurythmics for the title.

[†] Institut d'Anàlisi Econòmica (CSIC) and CREA. Email: sjaak.hurkens@iae.csic.es; Web: http://hurkens.iae-csic.org; Address: Campus UAB, 08193 Bellaterra (Barcelona), Spain.

[‡] University of California, San Diego. Email: nkartik@ucsd.edu; Web: http://econ.ucsd.edu/∼nkartik; Address: Department of Economics, 9500 Gilman Drive, La Jolla, CA 92093-0508.

1 Introduction

It is by now fairly well-established that people are motivated not just by material self-interest, but also by "social" goals (e.g. Fehr and Schmidt, 1999; Andreoni and Miller, 2002). Also important is that a person's preferences may have a procedural component: *how* allocations come to be can matter above and beyond just *what* the allocations are (Sen, 1997). That is, process matters beyond consequences. A prominent example is Gneezy's (2005) experimental study of aversion to lying. Gneezy compares people's behave in two different settings: a deception game where a person can tell the truth and obtain a "socially-minded allocation" (i.e, one that is more equitable or generous to the subject he is matched with) or lie and obtain a "selfish allocation"; or an otherwise identical dictator game where a person simply chooses between the two allocations.

Gneezy finds that a significant fraction of people display an aversion to lying or deception: the fraction of subjects who choose the selfish allocation in the dictator game is significantly higher than the fraction who make the same choice in the deception game by lying. We fully agree with this conclusion, and believe it to be important. By varying the monetary payments associated with the two allocations, Gneezy (p. 385) derives the following "main empirical finding."

Gneezy's Main Result. "*People not only care about their own gain from lying; they also are sensitive to the harm that lying may cause the other side.*" (p. 385) In the experiments, fewer people lie when the monetary loss from lying is higher for their partner, but the monetary gain remains the same for them. Similarly, fewer people lie when their own monetary gain decreases, while the loss for their partner remains the same.

This paper's first contribution is to show that the above empirical observations may be due entirely due to people's social preferences over different allocations, and not because of how lying aversion varies with allocations. In particular, we will demonstrate that Gneezy's data is consistent with the following hypothesis:

Hypothesis. *Conditional on preferring the outcome from lying over the outcome from truth-telling, a person is sensitive to neither her own [monetary] gain from lying, nor how much [monetary] harm she causes the other side.*

Notice that aside from the preface of "preferring the outcome from lying", the rest of the Hypothesis seems quite at odds with Gneezy's Result. The reconciliation is that a significant fraction of Gneezy's subjects prefer the outcome from lying to truth-telling in some experimental treatment(s), but not in others. But this by itself is merely evidence of social preferences, which as noted earlier, is a well-established phenomenon. It has nothing to do inherently with lying aversion, since it applies just as well when the allocations are chosen directly in the dictator game rather than through communication in the deception game. This highlights our analytical point: it is important to condition on preferences over allocations when interpreting lying behavior. Indeed, we will show that Gneezy's data is consistent with an even stronger version of the Hypothesis, viz., one cannot reject that 50 percent of people lie whenever they prefer the outcome from lying versus truth-telling, and 50 percent of people never lie.

That the Hypothesis is consistent with Gneezy's data does not imply that it is an accurate description of people's behavior. It is an important hypothesis to test because if it is right, it means that people can be categorized as one of two types: either they are "ethical" and never lie, or they are "economic" and lie whenever they prefer the allocation obtained by lying. If it is wrong, then a richer model of aversion to lying is needed.[1]

Accordingly, we ran a similar set of experiments to Gneezy's with the primary objective of testing the Hypothesis. Our secondary objective was to test the robustness of his findings, including how deception occurs in different cultures: his experiments were in Israel, we ran ours in Spain. Following Gneezy's design, we ran treatments of both the dictator game and deception game, but used a within-subject design so that players played both games (unlike Gneezy, where players played only one or the other); this permits us to make more precise inferences regarding people's decisions to lie relative to their preferences over allocations. We also used treatments that are more polarized than Gneezy's in terms of how much the dictator or sender can gain by implementing one allocation over another. This, in principle, should make it more likely to reject the Hypothesis, if the Hypothesis is wrong, while it should have no effect if the Hypothesis is correct. In this sense, our design reduces the possibility of type II errors. Further details of our design are postponed to Section 4.

Our data confirm Gneezy's finding that there is a statistically significant level of lying aversion. However, with regards to how this aversion to lying varies with consequences, even our data cannot reject the notion that so long as a person prefers the outcome from lying, the decision to lie is independent of how much she gains and how much her partner loses, i.e. we are unable to reject the Hypothesis.

1.1 Related Literature

Recently, there has been a growing interest in studying lying aversion experimentally. After completing our experiments, we became aware of the work of Sutter (2007), who documents in follow-up experiments to Gneezy that many subjects may not expect their lies to be believed. We also find this in our data. The focus of our papers are distinct, however. Sutter's (2007) main

[1] For example, Charness and Dufwenberg (2005) propose an application of "guilt aversion" theory to the current context.

concern is whether people who tell the truth in fact intend to deceive their partner; he does not run the control dictator games as Gneezy did and we do, and hence cannot speak to people's preferences over final outcomes.

Sánchez-Pagés and Vorsatz (2007) suggest that truth-telling preferences can explain the behavior they find in a sender-receiver game. While our focus here is on the interaction of lying aversion and social preferences of the sender, they study the receiver's preferences to punish the sender for lying, and how this affects lying behavior. In our experiment (as in Gneezy's), the receiver is not aware of the sender's incentives, and is never told whether the sender lies or not.

Finally, we mention three other studies. Rode (2006) investigates whether "context" affects senders' and receivers' behavior, and finds that receivers may be more trusting if the context is one of competition rather than of cooperation. Wang, Spezio and Camerer (2008) implement eyetracking and pupil dilation measurements in a discrete version of a Crawford and Sobel (1982) game; they suggest that their results indicate cognitive difficulties for senders in figuring out "how to lie." Ederer and Fehr (2008) study deception in a principal-agent tournament setting, also finding support for some lying aversion.

The remainder of this paper proceeds as follows. Section 2 describes Gneezy's (2005) experiments and findings in more detail. Section 3 develops our analytical points about the interpretation of his data. Our new experiments and findings are presented in Section 4. Section 5 concludes.

2 The Gneezy Experiments

Gneezy (2005) runs three treatments of a two-player experiment where there are only two possible outcomes, A_i and B_i, in each treatment $i = 1, 2, 3$. Although the actual choice between the options was to be made by player two, only player one was informed about the monetary consequences of each option. The only information player two had about the payoffs prior to making her choice was the message that player one decided to send. This message could either be "Option A_i will earn you more money than option B_i" or "Option B_i will earn you more money than option A_i".

In all three treatments, option A_i gives a lower monetary payoff to player one and a higher monetary payoff to player two than option B_i. (It is important to emphasize that player two did not know that.) Therefore, sending the second message can be considered as telling a lie, whereas sending the first message can be considered as telling the truth. The different monetary allocations (in dollars) in the three treatments were as follows, where as usual, a pair (x, y) indicates that player one would receive x and player two would receive y:

$A_1 = (5, 6)$ and $B_1 = (6, 5)$;
$A_2 = (5, 15)$ and $B_2 = (6, 5)$;

$A_3 = (5, 15)$ and $B_3 = (15, 5)$.

A fundamental issue when thinking about whether a player one would lie or tell the truth is what beliefs she holds about her partner's responses to her messages. Gneezy provides evidence suggesting that people generally expect their recommendations to be followed, i.e. they expect their partner to choose the option that they say will earn the partner more money. In this sense, lies are expected to work. While we return to the issue of beliefs in Section 4, we are content for now to accept Gneezy's interpretation about his subjects' beliefs, and will follow him in analyzing the situation as effectively a decision-theoretic problem for subjects in the role of player one.

In order to determine the extent to which the results of these deception games reflect an aversion to lying as opposed to preferences over monetary distributions, Gneezy used a control dictator game in which player one chooses between two options and where player two has no choice. Again, three treatments were run, corresponding to exactly the same options of the three treatments of the deception games.[2]

Each treatment of the deception game was run with 75 pairs of subjects. Each treatment of the dictator game was run with 50 pairs (consisting of different subjects from the deception game). We summarize the results in the following table, whose content is a replica of Table 2 from Gneezy.

Table 1—The percentage of player 1's who chose option B

Treatment	1	2	3
Deception	0.36	0.17	0.52
Dictator	0.66	0.42	0.90

The differences between the proportions in the Deception row are statistically significant (at the level of $p = 0.024$). Similarly, the differences between the proportions in the Dictator row are statistically significant (at the level of $p < 0.01$). Finally, for each treatment $i = 1, 2, 3$ the difference between the proportions of subjects choosing option B_i in the deception game and the dictator game is statistically significant (at the level $p = 0.01$). Gneezy concludes from this last point that people's choices reflect non-consequentialist preferences, since they treat the choice between A_i and B_i differently depending on whether it is an "innocent" choice or whether a lie has to be used to obtain it. We fully agree with this conclusion. In fact, when pooling over all three

[2] In order to make the comparison between the two games as fair as possible, it was announced to player one that his chosen option would be implemented with probability 0.8, while the other option would be implemented with probability 0.2. The reason for this was that in the deception game about 80 percent of the subjects in the role of player two follow the "recommendation" of player one, and that this was anticipated (on average) by the subjects in the role of player one. Note that Gneezy could do this since he ran the dictator game only after obtaining the results of the deception experiments.

treatments one finds an even much higher significance level for the difference in proportions of lies and innocent choices (in a Chi-square test, $X^2 = 33.21$, df$= 1$, $p = 0.0000$).

Gneezy (p. 385) asserts his "main empirical finding" to be that "people not only care about their own gain from lying; they also are sensitive to the harm that lying may cause the other side." This conclusion is drawn by comparing the percentage of liars across the three deception game treatments. Our primary concern is to what extent this conclusion is warranted. In the following section, we argue that his data cannot reject a model in which some fraction (e.g. half) of the population will say anything—be it the truth or a lie—to obtain their preferred outcome, whereas the remainder (e.g. the other half) are always honest. This implies that Gneezy's conclusion is only warranted to the extent that people's social preferences influence whether they actually prefer the outcome from lying relative to truth-telling, independent of any aversion to lying. Conditional on preferring the outcome from lying, a person may be completely insensitive to how much he gains or how much his partner loses from the lie.

3 Conditional Probabilities of Lying

The dictator control games show clearly that many subjects do not choose based only on their own monetary payoff; instead, many people take into account their partner's monetary payoff. In particular, out of 150 dictators, only 66 percent chose the option that gave them the highest monetary payoff; more than one third of the dictators chose to be *generous*, by which we mean choosing the option that gives the partner the highest monetary payoff. By revealed preference, a generous dictator prefers option A over B, although option B yields her a strictly higher monetary payoff. Surely, a person who prefers A over B will not tell a lie in order to obtain the less preferred option, B.[3] Only those who can shift the outcome in their preferred direction by lying need deliberate whether to lie or not.

Unfortunately, we do not observe which or how many of the subjects prefer option B over A (independently of lying) in the deception game treatments. This is because there are no subjects who played both dictator and deception games, as the experimental design was "between subjects", not "within subjects". However, the control dictator games corresponding to treatment $i = 1, 2, 3$ do give us an estimate of the percentage of people in the population who prefer option B_i over A_i, which we call *selfish* behavior. Let q_i denote the percentage of selfish people in treatment i, and p_i denote the fraction of liars in treatment i. Assuming that the subjects for each treatment of either game were drawn randomly from the same population distribution, q_i is then an estimate for the fraction of subjects who have any incentive to lie at all in deception game treatment i,

[3] Two caveats should be emphasized: first, recall that following Gneezy, we are assuming that messages will be followed by player two's; second, we assume that subjects do not derive inherent pleasure from lying.

and p_i represents what fraction actually do lie. In each treatment i, the ratio p_i/q_i is therefore an estimate of the fraction of people *who lie conditional on having an incentive to do so*, or for short, the *conditional probability of lying.* If the Hypothesis in the introduction is correct, then there should be no significant difference across treatments of this ratio. Even more strongly, if the conditional probability of lying is not statistically different from one half in any treatment, then one cannot reject the hypothesis that 50 percent of subjects are "ethical" (never lie) and 50 percent are "economic" (lie whenever they prefer the outcome from lying).

In treatment 1, 66 percent of the subjects revealed a preference for B_1 over A_1 in the dictator game. In the deception game, 36 percent of the subjects lied; hence, the fraction of people who lie conditional on having an incentive to do so, is about 55 percent ($\approx 36/66$). Doing similar calculations for the other treatments leads to the results that are summarized in Table 2.[4]

Table 2—The estimated conditional probabilities of lying when having an incentive to do so.

Treatment	1	2	3
Conditional probability	0.545	0.413	0.578

Table 2 clearly suggests that there does not seem to be a significant difference in the conditional probability of lying between treatments 1 and 3. Moreover, the difference in conditional probability of lying between treatments 2 and 3 is far less stark than the difference in absolute probability (cf. Table 1). In fact, it is straightforward to verify that none of the conditional probability differences are significant at the $p = 0.10$ level. We use the normal approximation to the binomial distribution to calculate p-values from a one-tailed test of the equality of the conditional probabilities of lying. The p-value for the comparison between treatments 2 and 3 equals $p = 0.104$. For the comparison between treatments 1 and 2 it equals $p = 0.170$. Finally, the comparison between treatments 1 and 3 yields a p-value of $p = 0.380$. (Appendix 1 provides detailed calculations.) We conclude that at the 10 percent level, one cannot reject the hypothesis that the conditional probabilities of lying in treatment 1 is *no different* from the conditional probability of lying in treatments 2 and 3. It is important to emphasize that our test takes appropriately into account that the number of subjects who have an incentive to lie in each treatment is a random variable.[5] In fact, for none

[4] It is worth noting again that for the purpose of this analysis, we follow Gneezy and assume that all subjects in the role of player one in the deception game expect their messages to be followed by player two. If some subjects have different expectations, it would affect the conditional probability level in any treatment. However, one would expect it to affect all treatments similarly, since expectations should be the same across treatments given that player two is told nothing about the possible payoffs.

[5] For example, ignoring integer problems, it would be a mistake to assume that in treatment 2, exactly 31.5 subjects have an incentive to lie (42 percent of 75), and in treatment 3 exactly 67.5 subjects (90 percent of 75) have an incentive to lie. Under this erroneous assumption one would then test for the equality of proportions of 13/31.5

of the treatments can one reject the hypothesis that this conditional probability equals one half.[6] Thus, one cannot reject either of the following: (i) 50 percent of people never lie and 50 percent lie whenever they prefer the outcome from lying; (ii) a person who prefers the outcome from lying flips a fair coin to decide whether to lie or not.

We would like to note that the magnitudes of the conditional probabilities in Table 2 do suggest that the Hypothesis may be incorrect: given that a person has an incentive to lie, the person is more likely to do so when her own monetary gain is bigger and when the monetary harm caused to the opponent is smaller. In this sense, it is suggestive that Gneezy's main result may in fact be correct. The problem is simply that the differences in these conditional probability estimates are not statistically significant given the sample sizes.

4 New Data

In an attempt to test the Hypothesis more carefully, we ran a set of new experiments; to permit comparison to Gneezy (2005), we retained the basic tenets of his design. The experiments were run over three sessions at the Universitat Autònoma de Barcelona in Spain; subjects were college students from various disciplines. No subject was allowed to participate in more than one session. Our subjects were given written instructions in Spanish; Appendices 2 and 3 provide English translations. There are five important differences in our design with respect to Gneezy's.

First, we had all subjects play both the deception game and the dictator game, unlike in Gneezy's experiment, where subjects played only one or the other game. In our design, both games are played with the same set of monetary payoffs, but each player is matched with a different, anonymous partner for each. (We paid subjects for only one of the games, determined by the flip of a coin after all decisions were made by all subjects—thus there is no feedback.) The reason we chose this

and 39/67.5, and one would find in a one-tailed test $p = 0.063 < 0.104$. But in fact, although in expectation 31.5 subjects have an incentive to lie in treatment 2, the probability that exactly n subjects have such incentives is equal to $\binom{75}{n} 0.42^n \times 0.58^{75-n}$, and similarly in treatment 3.

[6] Instead of using three pairwise comparisons, one can also test directly whether the conditional probability of lying is the same in all three treatments. Assuming that the true probabilities of having incentives to lie are given by the estimates from the dictator games, one can perform a Chi-square goodness-of-fit test by comparing observed frequencies of lies with expected frequencies of lies, given the null-hypothesis of equal conditional probabilities of lying. This gives $X^2 = 1.697$ with df$=2$ and $p = 0.43$, which means we cannot reject the hypothesis that all conditional probabilities are the same. Similarly, we can test whether all conditional probabilities are equal to one half. This yields $X^2 = 2.398$ with df$=2$ and $p = 0.30$. Again, we cannot reject the null-hypothesis of all conditional probabilities being equal to one half. An alternative way of analyzing Gneezy's data is to run a regression of the fraction of lies or selfish choices on (i) the difference between treatments, (ii) the difference between dictator and deception game, and (iii) the difference (between two treatments) in differences (between dictator and deception game). It turns out that the coefficient of "difference in differences" is insignificant, even at a 20 percent level (see Appendix 1). This is the analog of conditional probabilities of lying being constant over different treatments.

within subject design is that it allows us to directly compare any subject's behavior in the deception game with her preference over allocations as revealed by her choice in the dictator game.[7] With Gneezy's design, such comparisons can only be done at the aggregate level over all subjects, as we have done in the previous section. But this necessarily adds some noise to the estimates and leads to higher p-values, and thus to more type II errors of not rejecting the Hypothesis when it is in fact wrong. (See also fn. 5.)

Second, Gneezy adapted the dictator game so that the choice of player one was only implemented with probability 0.8. He did this to make it more comparable to the deception game. However, he was able to do this since he ran the control dictator games only after concluding the deception experiments. Since we used the within subject design, we could not make an adequate adaption so we implemented the choice of player one in the dictator game with probability 1.

Third, we conducted the experiment using the *strategy method* (Selten, 1967) for player two (receiver) in the deception game: rather than telling them what the message sent by player one (sender) is and asking them to pick an option based on it, we asked them to indicate which option they would pick contingent on each of the two possible messages from player one. The reason we chose this approach is that it allows us to directly observe a receiver's strategy. In particular, our design can identify the subjects who choose to ignore the message altogether—something that is impossible to detect using the *direct response method* employed by Gneezy.[8]

Fourth, we asked *all* subjects in the role of player one (sender) in the deception game to indicate their beliefs about what their anonymous partner would do in response to the message they chose. It should be remarked that we did not pay subjects for accuracy of beliefs. [9]

Finally, we conducted two different treatments, which we label 4 and 5 to preserve comparison with Gneezy's three treatments. In our treatments, the monetary payoffs, in Euros, were as follows:

$A_4 = (4, 12)$ and $B_4 = (5, 4)$;
$A_5 = (4, 5)$ and $B_5 = (12, 4)$.

[7] One potential concern with our design is that of order effects: does a player's behavior change depending on whether she plays the dictator game or deception game first? To account for this, we randomized subjects to play in both orders—deception game first or dictator game first—and found no significant order effects.

[8] The evidence on whether subjects behave differently in experiments that use the strategy method versus the direct response method is mixed. For example, Brandts and Charness (2000) find no difference in behavior across the two methods in simple 2x2 complete information sequential games, whereas Brandts and Charness (2003) do find some differences in an experiment involving communication about intentions and costly retributions.

[9] One practical reason for not paying for accuracy was that we ran the experiment using pen and paper. Calculating accuracies and exact payments would have taken too much time so that we would not have been able to pay the subjects immediately after the session. Moreover, we were mainly interested in learning whether the *average* expectation that recommendations would be followed was comparable with that reported in Gneezy's experiment. Gächter and Renner (2006) suggest that paying subjects for accuracy may lead to more *accurate* reporting but not necessarily to better *average* predictions.

Treatment 4 is similar to Gneezy's treatment 2 in the sense that option B entails a small gain for player one and a big loss for player two, relative to option A. Treatment 5 is substantially distinct from any of the three treatments in Gneezy because option B results in a big gain for player one and only a small loss for player two. If lying induces outcome B whereas telling the truth induces outcome A (as is suggested by Gneezy's data), and if the decision whether to lie or not depends on the relative gains and losses even conditional on preferring the outcome from lying, then one would expect to find that the proportion of lies among the selfish subjects in treatment 5 is significantly higher than in treatment 4. In other words, by using two treatments that are very polarized we increase the chance of rejecting the Hypothesis whenever the Hypothesis is wrong. That is, this aspect of our design further reduces type II errors.

Due to our within subject design, subjects in the role of player one can be divided into four categories based upon their preferences (selfish or generous) and their message (lie or truth). For example, a subject who chooses B in the dictator game but sends the message that A earns the receiver more money than B is classified as "Selfish and Truth". Table 3 reports the observed frequencies of the four possible types in each treatment.

Table 3—The number of liars and selfish subjects

Treatment	4	5
Selfish and Liar	19	14
Selfish and Truth	25	16
Generous and Liar	3	1
Generous and Truth	11	1
Total	58	32

As expected, the proportion of selfish subjects in treatment 5 ($30/32 \approx 0.94$) is significantly higher than in treatment 4 ($44/58 \approx 0.76$). (A one sided test of equality of proportions yields $p = 0.02$.) It is noteworthy that despite the similarity of treatments 2 and 4, the proportion of selfish subjects in treatment 4 ($44/58 \approx 0.76$) is significantly higher than in treatment 2 ($21/50 = 0.42$). (A one sided test of equality of proportions yields $p < 0.001$.) The data also show that the proportion of lies in the deception game ($41/90 \approx 0.46$) is significantly lower than the proportion of selfish choices ($74/90 \approx 0.82$). (A one sided test of equality of proportions yields $p < 0.001$.) These observations confirm Gneezy's finding of the existence of lying aversion.

We now come to the issue of whether the conditional probabilities of lying differ between the two treatments. The percentage of liars in treatments 4 and 5 are 38 percent (22/58) and 47 percent (15/32) respectively. This difference is not statistically significant: testing the hypothesis of equal proportions of liars versus the alternative hypothesis of a lower proportion of liars in treatment 4 yields a p-value of $p = 0.20$. When focussing only on the subset of players who are selfish, we find that the fractions of liars are 43 percent (19/44) and 47 percent (14/30), respectively. These percentages are not significantly different either: a one-sided test of equal proportions yields a p-value of $p = 0.38$.[10] These percentages are not significantly different from 50 either, so we must conclude that one still cannot reject the main Hypothesis we set out to test.

This came as a surprise to us. Based on Gneezy's data, we anticipated that our subject pool was big enough to yield significant differences in the fraction of selfish people who lie in our two treatments. We now turn to analyzing why this was not the case. While one possibility is that the Hypothesis is simply an accurate description of lying aversion for our subjects, we argue that receivers in our subject pool are far less trusting of the sender's message, and therefore, the incentives to lie are much attenuated for our subjects relative to Gneezy's. In fact, senders seemed to anticipate this rather low level of trust.

Let us first discuss the behavior of receivers in our experiment. Since the information given to a subject in the role of receiver is identical in both treatments, we find it reasonable to pool the data from both treatments for this purpose. Based on the realization of random matches, 59 out of 90 subjects (66 percent) chose the option that player one (sender) said would give a higher monetary payoff to player two (receiver). The other 31 out of 90 subjects (34 percent) chose the other option. In this sense, only 66 percent of 90 subjects followed the "recommendation", compared to the 78 percent of 225 subjects reported in Gneezy's experiments. This is a statistically significant difference: the null hypothesis of equal proportions versus the alternative hypothesis of less "recommendation following" in our experiment yields a p-value of $p = 0.012$.

Let us now discuss the behavior of senders. An important question is whether subjects, in the role of the sender, had a good estimate of how the receivers would act. Selfish senders who foresee that their message may be inverted may prefer to send the true message indicating that option A is better for the responder, hoping that the distrustful responder responds by choosing option B. We asked subjects in the role of sender (player one) which option they thought the receiver would choose based upon the message actually chosen to send. They could choose between saying that they expected A to be chosen, B to be chosen, or they were unsure.[11] Table 4 reports the expectations of the senders.

[10] Note that, if one only had coinciding data that were generated from subjects playing either the deception game or the dictator game (but not both), one would have estimated that 50 percent of the selfish subjects would lie, in both treatments. Namely, in treatment 4 there would have been 44 selfish subjects, and 22 liars while in treatment 5 there would be 30 selfish subjects, and 15 liars. Our experimental design allowed us to observe some difference between treatments, but this difference is not statistically significant.

[11] We included this third "unsure" option in order to distinguish the subjects who are very confident that their recommendation will be followed or inverted from those subjects who are not very sure.

Table 4—Expectations of senders

Expected reply	Trust	Unsure	Invert
Treatment 4	27	11	20
Treatment 5	11	11	10

A sender is classified as expecting "Trust" if he expects the receiver to choose the option that he says will give the receiver the highest monetary payoff. A sender is classified as expecting "Invert" if he believes the receiver will choose the option not recommended. The remaining subjects are classified as "Unsure".

In the following we will assume that a sender who expects that his recommendation will be followed did also believe that the other recommendation would also be followed.[12] Similarly, we assume that a sender who expects that his recommendation will be inverted, did also believe that the other recommendation would also be inverted. We interpret an answer of "unsure" that the sender attaches equal probabilities to option A or B after any message. Under these assumptions the average sender would believe the recommendation to be followed in treatment 4 by 56 percent of the receivers and in treatment 5 by 52 percent. This difference is not statistically significant.[13] Pooling the two treatments, we obtain 54 percent. This is significantly lower than the expectation of 82 percent that the subjects in Gneezy reported.[14] Quite remarkably, the expectations of the senders and the actual responses are fairly close. (Namely, 54 versus 66 percent in our experiments, and 82 versus 78 percent in Gneezy. The percentages of expected and actual responses are not significantly different.) It appears that there is some population-specific but well-calibrated level of trust and expected trust across the two subject pools.

[12] One might worry that this is not justified. That is, a sender who reports that he expects the receiver to choose option A after the message "Option A earns you more money than option B" may believe that the sender will choose option A in any case, or he may believe that the sender would have chosen option B in case he had sent the message "Option B earns you more money than option A". In the first two sessions we ran, we did not ask senders to report their beliefs about the reaction of the sender to the message that was not actually sent. Given the observation that many receivers chose a constant strategy, we wondered whether senders did foresee such reactions. In the third session we did therefore include a question about the hypothetical reaction to the unsent message. It turned out only 1 out of 30 subjects expected his message to be ignored. All other subjects were consistent in the sense that they expected the same reaction (trust, invert, or unsure) for both messages. This justifies our classification methodology.

[13] Testing the null-hypothesis of equal proportions of "trust" versus the alternative hypothesis of less trust in treatment 4 yields a p-value of $p = 0.37$.

[14] The p-value for testing the null-hypothesis of equal proportions versus the alternative hypothesis of lower expectations of trust in our experiments yields $p < 0.001$.

5 Conclusion

Gneezy (2005) shows convincingly that not all people are willing to use a lie to obtain their favorite outcome. However, the claim that people are more likely to lie when they can gain more and the partner loses less conflates distributional preferences and aversion to lying. A major point of this article is that the relevant probabilities of lying are those that are conditioned on having an incentive to lie. Interpreting Gneezy's data in this light shows that the change in lying behavior as payoff distributions are varied can be explained entirely by preferences over material payoffs.

As it is important to know how people's decisions with respect to lying depends on the consequences even when conditioned on merely preferring the outcome from lying, we ran additional experiments in Spain with a design that allows for the conditioning, and that uses more polarized treatments than Gneezy's. Despite this, we do not find statistical significant support for Gneezy's claim. We found that our subjects in Spain are less willing to follow the recommendations they receive. Instead, recommendations are often ignored or even inverted. Our senders seem to be aware of this possibility that lies will often not be believed and thus not work, since the average reported expectation of trust is fairly close to the actual average level of trust. This raises the possibility that the Israeli and Spanish student pools are different in this respect. However, the differences could also be caused by the different experimental designs. In particular, our use of the strategy method for receivers might have made it relatively easy for them to ignore (hypothetical) messages. Also, since we elicited beliefs of senders without giving monetary incentives for accuracy, the reported expectations of trust must be treated with caution.

Other questions that deserve further investigation concern the behavior of generous subjects. Will a generous person lie to his partner when he expects not to be trusted? More broadly, is there a correlation between standard social preferences are deontological preferences against lying? Such issues present interesting avenues for future research.

References

Andreoni, James and John Miller, “Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism,” *Econometrica*, March 2002, *70* (2), 737–753.

Brandts, Jordi and Gary Charness, “Hot and Cold Decisions and Reciprocity in Experiments with Sequential Games,” *Experimental Economics*, 2000, *2* (3), 227–238.

_ **and** _ , “Truth or Consequences: an Experiment,” *Management Science*, January 2003, *49*, 116–130.

Charness, Gary and Martin Dufwenberg, “Deception: The Role of Guilt,” 2005. mimeo.

Crawford, Vincent and Joel Sobel, “Strategic Information Transmission,” *Econometrica*, August 1982, *50* (6), 1431–1451.

Ederer, Florian and Ernst Fehr, “Deception and Incentives - How Dishonesty Undermines Effort Provision,” 2008. mimeo, MIT and University of Zurich.

Fehr, Ernst and Klaus M. Schmidt, “A Theory Of Fairness, Competition, And Cooperation,” *The Quarterly Journal of Economics*, August 1999, *114* (3), 817–868.

Gächter, Simon and Elke Renner, “The Effects of (Incentivized) Belief Elicitation in Public Good Experiments,” Discussion Papers 2006-16, The Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham September 2006.

Gneezy, Uri, “Deception: The Role of Consequences,” *American Economic Review*, March 2005, *95* (1), 384–394.

Rode, Julian, “Truth and Trust in Communication: An experimental study of behavior under asymmetric information,” 2006. mimeo, Universitat Pompeu Fabra.

Sánchez-Pagés, Santiago and Marc Vorsatz, “An experimental study of truth-telling in a sender-receiver game,” *Games and Economic Behavior*, October 2007, *61* (1), 86–112.

Selten, Reinhard, “Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments,” in H. Sauermann, ed., *Beiträge zur experimentellen Wirtschaftsforschung*, Tubingen: J.C.B. Mohr, 1967.

Sen, Amartya, “Maximization and the Act of Choice,” *Econometrica*, July 1997, *65* (4), 745–780.

Sutter, Matthias, “Deception through telling the truth?! Experimental evidence from individuals and teams,” 2007. forthcoming in *Economic Journal.*

Wang, Joseph Tao-yi, Michael Spezio, and Colin F. Camerer, "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth-telling and Deception in Sender-Receiver Games," 2008. mimeo, National Taiwan University and Caltech.

APPENDIX 1

In this Appendix we give details of the calculations performed in Section 3.

Conditional probabilities. Formally, let p_i denote the probability that a subject in treatment i who actually prefers B_i over A_i will choose to lie. Assume that persons who have no incentive to lie will not do so. Finally, assume that the probability q_i of having an incentive to lie in treatment i is exactly equal to the estimate given by the data of the dictator games. Hence, $q_1 = 0.66$, $q_2 = 0.42$, $q_3 = 0.9$. Let $\bar{X}_i$ denote the frequency of subjects lying in the deception game in treatment i. Below, we use Φ to denote the cdf of a standard Normal (mean 0, variance 1) distribution.

For the comparison of treatment 2 versus 3, note that under the null hypothesis of equal conditional proportions, we have $p_2 = p_3 = \hat{p}_{23} = (13 + 39)/(75q_2 + 75q_3) = 52/99 \approx 0.525$. Under the null hypothesis, $\bar{X}_3 - \bar{X}_2$ would be approximately Normal with mean $\hat{p}_{23}(q_3 - q_2) = 0.252$ and variance $[\hat{p}_{23}q_3(1 - \hat{p}_{23}q_3) + \hat{p}_{23}q_2(1 - \hat{p}_{23}q_2)]/75 = 0.0056159$. Hence, $P(\bar{X}_3 - \bar{X}_2 > 26/75) \approx 1 - \Phi((0.347 - 0.252)/\sqrt{0.0056159}) = 1 - \Phi(1.263) = 0.104$. The p-value equals 0.104 and one cannot reject the null hypothesis at the ten percent level.

Treatment 1 versus 2: $\hat{p}_{12} = (27+13)/(75q_1+75q_2) = 40/81 \approx 0.494$. Under the null hypothesis, $\bar{X}_1 - \bar{X}_2$ would be approximately Normal with mean $\hat{p}_{12}(q_1 - q_2) = 16/135 = 0.118$ and variance $[\hat{p}_{12}q_1(1-\hat{p}_{12}q_1)+\hat{p}_{12}q_2(1-\hat{p}_{12}q_2)]/75 = 0.00512117$. Hence, $P(\bar{X}_1-\bar{X}_2 > 14/75) \approx 1-\Phi((0.186-0.118)/\sqrt{0.00512117}) = 1 - \Phi(0.950) = 0.170$. The p-value equals 0.170 and one cannot reject the null hypothesis at the ten percent level.

Treatment 1 versus 3: $\hat{p}_{13} = (27 + 39)/(75q_1 + 75q_3) = 22/39 \approx 0.564$. Under the null hypothesis, $\bar{X}_3 - \bar{X}_1$ would be approximately Normal with mean $\hat{p}_{13}(q_3 - q_1) = 0.135$ and variance $[\hat{p}_{13}q_1(1 - \hat{p}_{13}q_1) + \hat{p}_{13}q_3(1 - \hat{p}_{13}q_3)]/75 = 0.00644847$. Hence, $P(\bar{X}_3 - \bar{X}_1 > 12/75) \approx 1 - \Phi((0.16 - 0.135)/\sqrt{0.00644847}) = 1 - \Phi(0.311) = 0.380$. The p-value equals 0.380 and one cannot reject the null hypothesis at the ten percent level.

Difference in difference regression.
For the comparison of Treatments 1 and 2, we run a linear regression of the form

$$Y = a + b \text{ DEC} + c \text{ TR2} + d \text{ DEC*TR2},$$

where Y denotes the fraction of lies (in the deception game) or selfish B choices (in the dictator game), a is a constant, DEC is a dummy variable taking value 1 in case of the deception game, and $TR2$ is a dummy variable taking value 1 in case of Treatment 2.

The following table reports the result of this regression; the important point being that the coefficient on $DEC * TR2$ is not significant (even at a 60 percent level):

Variable	Coefficient	Standard error	t	$P > \|t\|$
Constant	$+0.660$	0.065	$+10.21$	0.000
DEC	-0.300	0.083	-3.59	0.000
TR2	-0.240	0.091	-2.62	0.009
DEC*TR2	$+0.053$	0.118	$+0.45$	0.652

For the comparison of Treatments 2 and 3, we run a linear regression of the form

$$Y = a + b \text{ DEC} + c \text{ TR3} + d \text{ DEC*TR3},$$

where Y denotes the fraction of lies (in the deception game) or selfish B choices (in the dictator game), a is a constant, DEC is a dummy variable taking value 1 in case of the deception game, and $TR3$ is a dummy variable taking value 1 in case of Treatment 3.

The following table reports the result of this regression; the important point being that the coefficient on $DEC * TR3$ is not significant (even at a 20 percent level):

Variable	Coefficient	Standard error	t	$P > \|t\|$
Constant	$+0.420$	0.060	$+6.86$	0.000
DEC	-0.247	0.079	-3.12	0.002
TR3	$+0.480$	0.087	$+5.54$	0.000
DEC*TR3	-0.133	0.112	-1.19	0.234