Do Lies Erode Trust?

Glynis Gawn[*] Robert Innes[†]

July 2016

Abstract

Does honesty promote trust and trustworthiness? We investigate how being lied to (versus told the truth) in a Gneezy (2005) deception game affects behavior in a subsequent trust game with different players. Using a design that controls for potential treatment effects on payoffs, mood and beliefs about the overall propensity for honesty in the experiment, we find that the specific experience of being lied to significantly erodes trust and trustworthiness.

Keywords: : deception, lying aversion, trust, experiment
JEL Classifications: D03

[*] Economics, U.C., Merced, email: ggawn@ucmerced.edu
[†] Economics, U.C., Merced, email: rinnes@ucmerced.edu

> "Those who have been lied to... are resentful, disappointed, and suspicious. They feel wronged; they are wary of new overtures Those who share (the perspective of the deceived)... are all too aware of the impact of discovered and suspected lies on trust and social cooperation." Sissela Bok (*Lying*, 1978).

1 Introduction

Lies are a common and frequent phenomenon in everyday life. DePaulo et al. (1996) found that college students and members of a U.S. community lied in 20 to 31 percent of social interactions recorded in daily diaries, with college students telling an average of two lies per day and community members one lie per day. In a recent survey of over 23,000 high school students, 76 percent self-reported that they had lied about something significant in the past year, while 38 percent indicated that they sometimes lie to save money (Josephson Institute of Ethics, 2012).

Such statistics would seem to be consistent with the standard economic model of self-interested behavior that predicts lying whenever an individual can materially benefit from this behavior. However, a recent economics literature provides compelling evidence that many individuals are averse to lies; they are honest despite monetary incentives to be dishonest (see, for example, Gneezy, 2005; Gibson et al., 2013; Fischbacher and Heusi, 2013; Abeler et al., 2014). While these results are a surprise from the standpoint of the baseline economic model, what if lies have serious economic consequences, beyond direct and immediate costs to the recipient of the lie? Perhaps lying aversion is a symptom of these consequences.

Two perspectives suggest that lies are likely to cause economic harm. On one hand, philosophers have elucidated the deleterious effects of lies, particularly on those who have been lied to, in works dating back as far as Aristotle (Bok, 1978). Lies, it is argued, erode trust and thereby deter social cooperation. On the other hand, recent economics research identifies significant benefits of generalized trust in promoting economic growth and progress.[1] If both perspectives are right, then values, norms and institutions that deter lies

[1] A compelling literature identifies close links between survey indicators of trust and, respectively, economic

– many of which we see in practice (Fischbacher and Heusi, 2013) – may deliver economic benefits by promoting trust.

In this paper, we study the effects of lies on those at the receiving end in order to test the general proposition that lies erode trust. Of course, not all lies are harmful. Our focus is on "black lies" – those that benefit the liar at the expense of the recipient and that thereby violate commonplace norms for honest conduct – as opposed to "white lies" that benefit the recipient (Erat and Gneezy, 2012). We measure how being lied to, in a first-round Gneezy (2005) deception game, alters behavior on both sides of a second-round trust relationship with a different person. Second round outcomes include whether to trust a partner and whether to reciprocate trust with trustworthiness in a simple interaction patterned after the original game of Berg et al. (1995).[2] We find that being lied to in a prior interaction erodes both trust and trustworthiness.

This conclusion is both general – in the sense that lies erode trust overall – and specific in the sense that lies erode trust even when controlling for a variety of correlated effects. What is it about the experience of a lie that might deter generalized trust? Is it because lies disappoint their recipients? Is it because lies "burn" their recipients by reducing payoffs they enjoy from the interaction? Is it because lies signal something about norms of honesty, so that a recipient of a lie reasonably infers that, generally speaking, others are less honest? Or is there something more fundamental about a lie that affects trust, separate from immediate disappointment, harm, and inferences about social behavior? For example, Bok (1978) stresses the fundamental nature of truthful communication as a cornerstone of human interaction; shaking this foundation with the experience of a lie, she suggests, can limit

growth (e.g., Knack and Keefer, 1997; Zak and Knack, 2001; Guiso et al., 2004), international trade (Guiso et al., 2009), development of financial institutions (Guiso et al., 2008), and other indicators of economic success (LaPorta, et al., 1997). A large experimental literature on trust games is arguably motivated by these links (see Johnson and Mislin, 2011, for a recent survey). Indeed, a growing body of work identifies the close relationship between behavior in experimental trust games and survey evidence on trust. See, for example, Glaeser et al. (2000), Lazzerini et al. (2004), Fehr et al. (2003), and Bellemare and Kroger (2007). This literature studies, among other things, the correlation between responses to survey questions on trust (such as answers to World Values Survey) and experimental indicators of trust and trustworthiness. Some of this work indicates correlation between survey measures and trustworthiness, but not trust (Glaeser et al., 2000, for example); others document correlation between survey responses and trust (Fehr et al., 2003, for example). Sapienza et al. (2013) reconcile conflicting evidence in their recent study.

[2] See also the lost wallet game of Dufwenberg and Gneezy (2000).

free will, jeopardize accumulation of knowledge, and, in the extreme, lead to the collapse of social institutions. These arguments suggest that a lie represents a poignant violation of social norms that is likely to have intrinsic consequences.

In this paper, we try to isolate an intrinsic effect of lies – the specific effect of experiencing a norm violation – separate from other forces. Changes in perceived social norms can spill over from one context to another (Keizer et al., 2008; Houser et al., 2012). In our setting, changes in perceived propensities for honesty could reduce trust in a subsequent interaction. We control for this channel of effect by informing all players in our experiment about the overall proportion of lies, so that the experience of being lied to, or told the truth, is an individual experience and not a reflection of norms or general propensities for honesty. Moods can also affect trust behavior (e.g., see Capra, 2004; Kirchsteiger et al., 2006), and a low payoff/"burn" can alter behavior due to effects on mood, preferences, and perceptions of procedural justice (Sanchez-Pages and Vorsatz, 2007, 2009). These general (payoff/mood) effects are symptoms of many experiences, including being at the receiving end of a low payoff from a dictator (e.g., Ben-Ner et al., 2004; Herne et al., 2013) or a first-round defector (Ellingsen et al., 2009) or, in our case, a lie. We seek to identify the experience effect of our treatments by controlling for payoffs, mood, and beliefs.

Individual experience has been found to drive economic behavior in other settings. For example, experiences in financial markets can affect stock investments and risk aversion (Malmendier and Nagel, 2011); experiences of inflation can affect inflation expectations and home ownership (Malmendier and Nagel, 2016); experiences of financial crises can alter managers' choices of external financing and leverage (Malmendier, Tate and Yan, 2011); experience of an insurance game can increase farmers' uptake of weather insurance products (Cai and Song, 2011). A larger literature studies learning and shows that individual experience can be over-weighted in behavioral choices (see, for example, Simonsohn et al., 2008; Camerer and Ho, 1999; Camerer, 2003). We find that individual experience of a norm violation can affect trust behavior, even though the experience has no relevant informational content.

Our results are relevant generally to the process of trust formation and transmission.

Butler et al. (2015) recently present experimental evidence that trust values (trustworthiness) are largely affected by parental upbringing, while trust beliefs (that support values) are largely affected by a false consensus – the belief that others act like oneself (Ross et al., 1977; Ellingsen et al., 2010). These mechanisms help to explain both heterogeneity in trust beliefs and persistence of this heterogeneity. An alternative (although not entirely competing) perspective embeds updating of beliefs based on experience and prevailing local norms (Guiso et al., 2008; Dohmen et al., 2012). Our results highlight the potential role of an individual's specific experience in driving both values (trustworthiness, given beliefs) and beliefs, and are consistent with a false consensus. Even though our treatments have no bearing on overall propensities for truthfulness, they drive choices and, consistent with false consensus effects, alter beliefs in the trust game.

These conclusions are optimistic in the sense that they suggest possible mechanisms to increase trust. For example, the adverse experience effects of lies that we find here can potentially be mitigated by traits, policies and norms that deter lies – including ingrained lying aversion, societal values that promote veracity, and a culture of honesty in organizations.

2 Relationship To Prior Literature

A large and growing economics literature studies what drives individuals to be honest despite monetary benefits from lying. Mostly drawing on Gneezy's (2005) initial deception game, recent studies have found that lying aversion varies across individuals (Gibson, Tanner and Wagner, 2013) and is sensitive to a variety of economic forces. These include the direct monetary consequences for those on both sides of the interaction (Gneezy, 2005; Gibson et al., 2013; Freeman and Gelber, 2010), strategic considerations (Sutter, 2009), social cues on how often others lie (Innes and Mitra, 2013), guilt aversion (Battigalli et al., 2013; Charness and Dufwenberg, 2010), gender (Dreber and Johanneson, 2008), the extent of the lie (Lundquist et al., 2009; Fischbacher and Heusi, 2013), and cooperation in prior play (Ellingsen et al., 2009) but not cooperative (vs. competitive) priming (Rode, 2010).[3]

[3] See also Mazar et al. (2008) on the role of self-concept and the recent survey by Rosenbaum et al. (2014).

In the present paper, we focus instead on the consequences of a lie, beyond its direct and immediate effect on payoffs to the liar and recipient of the lie, for trust. While there is a rich literature studying individual drivers of trust - including beliefs about behavior (e.g., Sapienza et al., 2013; Costa-Gomez et al., 2010), mood (e.g., Capra, 2004; Kirchsteiger et al., 2006), and a variety of preference attributes[4] – to our knowledge this is the first study investigating how the receipt of a lie affects trust interactions with different players.

We build on a handful of key studies that examine effects of lies on behavior. Gneezy et al. (2013) find that Receivers in a multi-round deception game are less likely to follow the recommendation of their Senders if they have been negatively affected by a lie in the previous round. While this result can be interpreted as a negative effect on trust, it may reflect learning from prior experience in the same (deception) game.

Several other papers consider effects of lies on subsequent interactions with the same player.[5] Tyler et al. (2006), uses videotaped conversations to reveal lying behavior to the participants. The authors find that when participants witness a partner lying, they like and believe the partner less and also increase their own use of deception in follow-up interactions with the same partner. Schweitzer et al. (2006) find that subjects who have been lied to twice in a row (with the second lie more egregious than the first) are less trusting of the liar relative to a player who has not communicated, perhaps due to learning and related beliefs about the partner's trustworthiness. Brandts and Charness (2003) show that deceptive (vs. truthful) messages lead to significantly more punishment when the Receiver obtains a low payoff as a result of a Sender decision; importantly, this conclusion controls for the payoff to the Receiver and, hence, does not reflect a "low payoff" effect.[6] Sánchez-Pagés

[4] Relevant attributes include altruism (Cox, et al., 2008; Ashraf, et al., 2006), reciprocity (Rabin, 1993; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004), inequity aversion (Fehr and Schmidt, 1999), risk aversion (Houser et al., 2010), values of social welfare (Charness and Rabin, 2002), benefits of a "warm glow" (Andreoni, 1990), and guilt aversion (Charness and Dufwenberg, 2006). See also Al-Ubaydli et al. (2013) who study how market priming promotes trust, indicating a causal connection between markets and the trust that is also associated with economic progress. (This is a small subset of the literature, and we apologize to authors of many key papers omitted here.)

[5] See also Duffy and Feltovich (2006) who consider the effect of learning about a prior lie (or truth) of one's partner (to someone else) on coordination games with that partner. They find that crossed signals worsen coordination relative to single signals.

[6] Brandts and Charness (2003) study a unique second stage game in which the Receiver has a dominant strategy and the Sender's choice determines whether this strategy produces payoffs that are (A1) better for

and Vorsatz (2007, 2009) examine the extent to which Receivers in a deception game punish their respective Senders. They find that the Receivers punish primarily when they have been lied to and been burned as a result (because they followed the lie), reflecting concerns for procedural justice.[7] While these interesting results come closest to identifying an experience effect of lies on the punishment behavior of Receivers, they reflect reciprocal responses to the player making the lie.

Our focus instead is on the generalized effects of receiving a lie, that is, how the experience affects attitudes and behavior in *an unrelated interaction with a different person.* This pins our study broadly in a literature on generalized indirect reciprocity (Stanca, 2009; Alexander, 1987). The literature distinguishes between direct reciprocity (when A takes an action that affects B, how does B reciprocate in an action that affects A?); social indirect reciprocity (how does another player C reciprocate in an action that affects the actor A?); and generalized indirect reciprocity or "paying it forward" (how does B reciprocate in an action that affects another player C?). The last phenomenon – our focus – has been studied in dictator games (Ben-Ner et al., 2004; Herne et al., 2013), trust games (Dufwenberg et al., 2001; Greiner and Levati, 2005) and gift exchange games (Stanca, 2009). Houser et al. (2012) study effects of a first-round dictator game on behavior in a subsequent (unrelated) cheating game with a different player, finding evidence of cross-context spillovers in social norms. A broad message from this literature is that generalized reciprocity is prevalent: a prior experience at the receiving end of perceived moral or immoral behavior affects subsequent behavior in a game also with moral overtones and a different player.

Studies of generalized indirect reciprocity embed a number of forces that can drive treatment effects on subsequent behavior. These forces include (1) the experience of payoffs and related effects on mood, and (2) inferences about norms and social behavior. We strive to pinpoint an experience effect, in the context of lies, that is distinct from these other forces.

the Receiver and worse for the Sender, or (A2) better for the Sender and worse for the Receiver. In the third stage, after knowing what happened in the second stage, the Receiver can reward or punish the Sender. In the first stage, the Sender chooses a non-binding message about his/her own decision (A1 or A2). The main result is that punishment increases under decision A2 when the message is false, A1, vs. true, A2.

[7] Sánchez-Pagés and Vorsatz (2009) introduce a costly silence option for Senders in a Gneezy (2005)-type game, showing how the presence of a punishment option promotes silence.

For example, Dufwenberg et al. (2001) compare trust games in which a Returner, when trusted by a Sender, alternately makes a decision on reciprocating by returning money to the same Sender (direct reciprocity) or to a different Sender (indirect reciprocity). Being trusted in these contexts may reflect a "good payoff" situation and signal a norm of trust and an expectation of trustworthiness.

We are interested generally in whether being at the receiving end of a norm violation relaxes one's own moral preferences.[8] For example, in the Andreoni and Bernheim (2009) model, experience may affect individual weights on norm compliance; experiencing a lie can affect the weight an individual places on a trustworthy (vs. selfish) action. Alternately, in a model of intention-based reciprocity (Rabin, 1993; Charness and Rabin, 2002) or guilt aversion (Charness and Dufwenberg, 2006; Battigalli and Dugwenberg, 2007, 2009) or spite (Levine, 1998), experience can alter the weight on reciprocation or a partner's disappointment or a superior allocation. Do lies have these effects?

3 The Experiment

Our design involves two subject interactions, in two games, between different players. First is the deception game, followed by the trust experiment. To avoid cross-game payoff spillovers, subjects are paid for only one of the two games, each with equal probability.

3.1 *The Deception Game*

The deception game follows the Gneezy (2005) design. In this game, Senders from one classroom are randomly paired with Receivers from another classroom, one Receiver for each Sender. The Sender observes two possible payoff allocations between the two players. In our game, the payoff options are as follows:

Option A: $6 to the Sender and $3 to the Receiver.

Option B: $4 to the Sender and $6 to the Receiver.

[8] In this sense, our results may add to a growing literature on what determines "moral wiggle room" (Dana et al., 2007; Charness & Sutter, 2012; Bartling & Fischbacher, 2012; Bartling et al., 2014; Grossman, 2015).

The Sender chooses one of two Messages to deliver to the Receiver, one truthful (Message B) and the other untruthful (Message A). The two possible Messages are:

> **Message A**: "Option A will earn you (the Receiver) more money than Option B."
>
> **Message B**: "Option B will earn you (the Receiver) more money than Option A."

Based only on the Message sent by the Sender, the Receiver chooses one of the two Options, which in turn determines payoffs to the two players, Sender and Receiver. In the experiment, Option labels are varied between subjects (sometimes Option A is better for the Receiver and sometimes Option B).

Following Gneezy (2005), Receivers are never told the dollar amounts in the two options. The intent is to avoid strategic conjectures by Receivers and strategic incentives for Senders (Gneezy, 2005, pp. 386-7). A lie can then be interpreted by a Receiver as a deceptive act by a Sender who anticipates a credulous response to the chosen Message (see Section 4 for further discussion on the interpretation of lies).

Sender sessions for the Gneezy (2005) game are conducted before the Receiver sessions (with different players in different classrooms). *Our focus is on the Receivers.* After all Receiver decisions are made in the Deception game, and the decisions collected by the experimenter, Receivers are exposed to our Treatments. Figure 1 depicts the timeline for the experimental sessions and the order of play for the Receivers. Figure 2 summarizes the design of the Receiver experiment, as described below, with game trees.

3.2 *The Treatments*

Each Receiver is randomly assigned to one of three Treatment groups. The first group is the set of Control subjects who are exposed only to common information about the Deception game, information that is given to all Receivers. The common information supplied in the treatments is as follows:

> "One of the two payment Options in Experiment 1 is BETTER FOR YOU, and the other Option is BETTER FOR YOUR SENDER."

> "In this session of Experiment 1, roughly 6 out of 10 Senders TOLD THE TRUTH and 4 out of 10 Senders LIED about the payment Option that earns Receivers (like you) more money."

The first statement conveys that a correct message is a "truth" and a false message is a "black lie" that is likely to be perceived as a norm violation (our intent).[9] The second statement is based on the Sender sessions of our experiment. Its purpose is to control for subject beliefs about behavior in the Deception experiment. The precise percentage of truthful Senders was 59.4 percent.

In addition to the common information described above, each subject in the second and third treatment groups is told whether his or her own matched Sender actually *lied* (treatment LT) or *told the truth* (treatment TT) in the Message that was sent. We are interested in the effects of this specific experience on subsequent decisions in a trust game. How does being lied to (versus being told the truth) affect one's willingness to trust and one's trustworthiness?

Random assignment to treatments is ensured by a random distribution of experimental questionnaires to student participants in each Receiver session. Each questionnaire identifies a participant by a registration number. Each registration number is associated with a particular Sender and a specific treatment group, with the correspondence known only by the experiment manager. If the matched Sender lied, the associated treatment group is either the LT or the Control. Similarly, if the Sender told the truth, the associated treatment group is either the TT or the Control.[10] Treatment assignments are made to obtain roughly

[9] In the treatments, we do not provide details on the precise dollar options in order to (1) follow the Gneezy (2005) design as closely as possible and (2) focus Receivers on the norm violation content of a lie (vs. truth), rather than the specific monetary / payoff consequences. This approach helps to reinforce the payoff independence between deception and trust games that is a key feature of the experiment (see related discussion in Section 4.2).

[10] Registration numbers contain a numerical identifier specific to each individual subject, followed by an alphabetical identifier associated with the treatment group (M, N, and P, for example). Alphabetical identifiers are different in different classrooms and known only to the experimenter. After turning in their

equal LT and TT sample sizes, and a slightly smaller Control group.[11] Specifically, out of 266 Sender matches in the experiment, 108 (30.4 percent) lied and 158 (59.4 percent) told the truth. Of the 108 lie matches, 91 Receiver questionnaires are assigned to the "Lied To" (LT) treatment and the remaining 17 are Controls. Similarly, of the 158 truth matches, 94 Receiver questionnaires are assigned to the "Told the Truth" (TT) treatment and the remaining 64 are Controls. Questionnaires associated with the three treatments are equally mixed for each classroom and are randomly distributed to student participants.

3.3 *The Trust Game (Experiment 2)*

After receiving the Treatments, subjects participate in a second experiment. Each participant is again matched with another player in a different classroom. None of the participants in this game are Senders from the Deception experiment, and subjects are told that their matched player is a different person than their Sender from Experiment 1.

Subjects are either in the role of Sender or Returner and each player starts with $4. The Sender chooses between two alternatives:

> **KEEP.** Keep the initial $4, implying that both players earn the $4 allocated to them.
>
> **SEND.** Send his/her $4 to the Returner.

deception game decisions, Receivers are given an information sheet. Control subjects (with M identifiers, for example) collect their sheet at one "station" to which they are directed (so that their information only reflects overall propensities for honesty of Senders). LT and TT treatment subjects (with N and P identifiers, for example) are each directed to another "station", where they are given an information sheet containing both information on overall propensities for honesty AND information on whether their own Sender lied or told the truth. This is the only point at which any reference is made to the alphabetical identifier. The correspondence between "station" numbers (1 and 2) is varied from classroom to classroom; sometimes station 1 is for the Controls and sometimes station 2. On the information sheet, the LT and TT treatment subjects are told (for the N and P example):

> "If your Registration number ends with an N, your Sender TOLD YOU THE TRUTH in Experiment 1 about the Option that earns you more money.
> If your Registration number ends with a P, your Sender LIED TO YOU in Experiment 1 about the Option that earns you more money."

To verify understanding, we also ask each of the LT and TT treatment subjects to circle whether they were Told the Truth or Lied To. All of our subjects answer this question correctly.

[11] The target size of the Control group is large enough to produce a mix of lies and truths in the Control, but slightly smaller than LT and TT counterparts (our main focus) in order to produce larger samples in the latter treatments.

If the Sender chooses SEND, the $4 sent becomes $8, which combined with the Returner's initial $4, makes $12 available. In this case, the Returner chooses between:

> **Option C.** Return $7 to the Sender, so that the Returner receives $5 and the Sender receives $7.
>
> **Option D.** Return $2 to the Sender, pay a fee of $2 and keep the remainder, so the Returner receives $8 and the Sender receives $2.

In this game, a SEND decision by the Sender is an indication of *trust*, and a Returner choice of Option C indicates *trustworthiness*. Table 1 summarizes the payments. Option labels are again varied between subjects (sometimes Option C is generous and sometimes stingy).

Table 1: Trust Game Payoffs

Returner's Option Choice	If Sender Chooses SEND		If Sender Chooses KEEP	
	Payment To Returner	Payment To Sender	Payment To Returner	Payment To Sender
C	$5	$7	$4	$4
D	$8	$2	$4	$4

In the experiment, participants make decisions in both roles. If Experiment 2 is selected for payment, a matched pair is paid according to one player's decision as Sender and the other player's decision as Returner; each of the two possible allocations of roles is implemented with equal (50 percent) probability. Subjects are told this procedure at the start of the trust experiment, with the corresponding instruction: "You should therefore make your decision in each situation (role) as if you will be paid according to that situation." Because participants simultaneously and anonymously make choices in both roles (Sender and Returner), with payments determined according to one of the two roles, reputational motivations are avoided.

Some aspects of our design might limit comparison to some other experiments. The use of a two-role protocol could potentially lead to different behavior than in experiments where subjects play only one role.[12] We also have participants make each type of decision only

[12] The literature gives a somewhat mixed picture on "role reversal" versus single role designs (Brandts and Charness, 2011). A number of authors have subjects play both roles in the trust game (for example, Chaudhuri and Gangadharan, 2007; Altmann et al., 2008). Charness and Rabin (2002), building on other literature, also have participants play both roles in a trust-type game that is played sequentially. In a

once, whereas many experiments have participants make the same decision repeatedly. We use the strategy method for the Returner (who answers contingent on a SEND decision), rather than a direct response approach.[13] We do not believe that these design choices are important factors in our results. What is important for our experiment is that the subjects' choices reflect "trusting" and "trustworthy" behaviors, an interpretation that is intrinsic to the standard trust game framework to which we adhere.

3.4 *Measuring Mood*

One possible mechanism by which specific experience (of being lied to, in our case) may affect behavior is due to its effect on a subject's mood. We ask subjects to gauge their mood at the start of the experiment (before instructions for the Deception game) and after completion of the treatments (but before the trust game), using the following scale:

bad down so-so good very good great

3.5 *Beliefs and Questions*

At the very end of the experiment, we use an incentive compatible approach to elicit beliefs about rates of trust and trustworthiness among students in the experiment. Subjects are asked to predict the fraction of participants who choose to Send and the fraction who choose to Return $7. Regardless of which game is selected for a subject's payment (deception or trust), each prediction is rewarded with a $1 payment if it is within 5 percent (plus or minus) of the true percentage (using 5 percentage point bands). In addition, we ask subjects to report their gender and to answer four questions indicating their interpretation of a lie vs. truth, as described below. A full set of experimental instructions can be found in the on-line Appendix.

subsequent paper, Charness and Rabin (2005) find that playing two roles (versus one) has no significant impact on their earlier results. Burks et al. (2003) study effects of two-role versus one (direct) role play in a trust game, when players are paid in both roles; they find that when participants are informed a priori that they will play both roles, there is a tendency to be less trusting and less trustworthy. These results suggest that the two-role design may potentially improve subjects' understanding of the game.

[13] Brandts and Charness (2011) provide evidence that the strategy method generally does not elicit significantly different responses in a variety of games, including trust.

4 Interpretation of Treatments and Hypotheses

4.1 *On the Meaning of Lies*

Does a message sent in the Gneezy (2005) game represent a lie or a truth, as we presume in the experiment?[14] The standard definition of a lie contains four ingredients (see Mahon (2016) for an exhaustive philosophical exposition on this subject): A lie is (1) a statement made by one who (2) believes the statement to be false, to (3) another person (4) with the intention that the other person believe that statement to be true. Add to this the simplest definition of deception (Mahon, 2016): to cause to believe what is false. A Gneezy (2005) lie satisfies all four criteria.[15] From the standpoint of the Receiver, the deception clause on the actual effect of a lie vs. a truth is germane: a lie causes belief of the untruthful message. The latter can be important in interpreting our experiment, as Receivers who follow their Sender recommendations presumably believe the message sent.

A potential worry is that subjects do not perceive the game in a normative light and instead act as economists would normally predict – as self-interested Nash players. If subjects assume conflicting incentives in the Gneezy (2005) game, then the unique Nash equilibrium is for Senders and Receivers to play each of their two respective options with equal probability (the cheap talk equilibrium, as described in Sanchez-Pages and Vorsatz, 2007). Under Nash outcomes, our treatments should convey nothing but the realization of chance.

On one hand, this conjecture presents the empirical challenge for our study. If subjects play Nash, then the information (and experience) provided in our treatments should have no effect. On the other hand, there are two ways to judge the conjecture's relevance to our experiment.

First, we directly gauge whether our Receivers interpret the Gneezy (2005) outcomes

[14] We thank the Editor, Professor Aoyagi, and an anonymous referee for raising the issues addressed in this Section.

[15] Sutter (2009) observes that many Senders predict that their Receivers will not follow their messages, giving rise to sophisticated liars (reject-predicting truth-tellers) and benevolent liars (reject-predicting liars). However, the majority of Senders in our experiment (like others) predict that their Receiver will follow their message, 69.3 percent on average in our case (76.6 percent for liars and 63.9 percent for truth tellers). While intentions are impossible to know, these responses suggest an intention to deceive (or not) the Receiver.

as lies vs. truths – as norm violations vs. norm compliance – by asking them to agree or disagree with each of the following statements: "Sending a false message is:

a) not really lying, just being rational.

b) trying to deceive/trick the Receiver.

c) not the right thing to do."

To judge the perception of the last statement (c) as a norm, we also ask subjects to predict the proportion of participants who agree that a false message is "not the right thing to do." The questions are all posed at the very end of the experiment, after all choices have been made. The Receivers' answers broadly indicate a perception that false messages are norm violations/lies. [16]

Second, there is substantial evidence that subjects in classrooms, labs and the field do not in fact behave according to Nash. A significant fraction are strictly averse to lies of the type we model, producing many fewer lies and more trusting behavior (by Receivers) than theory would predict.[17] Our experiments are no exception. We find that almost 75 percent of our Receivers follow their Sender message (and 78 percent did so in Gneezy's (2005) experiments), significantly higher than the Nash prediction of 50 percent. [18]

Such outcomes are consistent with an equilibrium in a simple model of a Gneezy-type game provided in the Appendix. In the model, a fraction of Senders are lie averse and Receivers have one of two beliefs about this fraction, one above fifty percent (a high belief) and one below (a low belief). The majority have the high belief.[19] In the equilibrium: lie-

[16] See Section 6.2 and Table 7 for details.

[17] See Sanchez-Pages and Vorsatz (2007) for an explicit test of Nash behavior in a Gneezy (2005) type game.

[18] The z statistic for the null of a 50 percent Receiver follow rate in our experiment is:

$$z = \frac{\overline{p} - 0.5}{\sqrt{[\overline{p}(1-\overline{p})]/N}} = 9.136$$

where $\overline{p}$ = sample mean Receiver follow rate = 0.744, N = number of observations = 266, and z is approximately standard Normal under the null. The corresponding p-value is less than 0.001.

[19] The model assumes that (i) Senders are either "high lie-averse" or zero-lie-averse, (ii) Receivers know the two possible Sender types and understand that Senders have conflicting incentives, and (iii) the distribution of Receiver beliefs is public information. If the majority of Receivers have the low belief, there is a unique mixed strategy Nash Equilibrium in which Receivers accept with 50 percent probability on average and zero lie averse Senders tell the truth with a probability less than one half. The modeled game differs from our Gneezy (2005) experiment because we only tell Receivers about the conflicting incentives after the deception

averse Senders tell the truth; zero-lie-averse Senders lie; high-belief Receivers expect truths with high probability, therefore accept their Sender recommendations, and are surprised by lies; and low belief Receivers expect lies, therefore reject, and are surprised by both truths and social information indicating a majority of truthful Senders.

4.2 *Hypotheses*

Our focus is the impact of the *Lied To* versus *Told the Truth* experience on trust decisions. We measure trust by the fraction of subjects who decide to Send (vs. Keep) in Experiment 2, and trustworthiness by the fraction of subjects who choose to Return \$7 (vs. Return \$2) in the event that their matched Sender chooses Send. The overarching hypothesis is:

> *Hypothesis 1 (H1).* The *Lied To* treatment elicits less trust and less trustworthiness than does the *Told the Truth* treatment.

To evaluate Hypothesis 1, our design controls for treatment effects on beliefs about social behavior, mood, and payoffs. We control for beliefs by informing all subjects about the proportions of Senders in the deception experiment who are truthful and untruthful. We explicitly measure mood and treatment-induced mood changes in order to disentangle any mood component of the treatments. We randomize between the two games (deception and trust), so that outcomes in the trust game are only implemented if outcomes from the deception game are not. This is a standard approach to eliminating cross-game payoff spillovers, including strategic hedging motives.[20]

Despite these design controls, the content of our treatments varies along two dimensions: whether the experience represents (1) a norm violation (a lie) vs. a norm compliance (a truth), and (2) an unanticipated vs. anticipated outcome. For first round accepters – who follow the Sender's Option recommendation in the deception game – the presumption is that the message is truthful; a lie is unanticipated and a truth is anticipated. The opposite is true for first round rejecters who do not follow their Sender's Option recommendation. Table 2

experiment is finished.

[20] See, for example, Hurkens and Kartik (2009), Altmann et al. (2008), Brandts and Charness (2011), and many others.

breaks down the treatment experience along the two dimensions. Our main interest is to learn how the norm violation component of the experience affects behavior.

Table 2: Content of Treatment Experience

		Norm Violation Vs. Compliance	
		Norm Compliance	Norm Violation
Anticipation Of Outcome	Anticipated	TT-A	LT-R
	Not Anticipated	TT-R	LT-A

TT = Told the Truth, LT = Lied To, A = first-round accepter, R = first-round rejecter

Why might the two dimensions be important? First, the primary psychological mechanism for experience effects is the availability heuristic (Nesbitt and Ross, 1980; Tversky and Kahneman, 1974; Hertwig et al., 2004; Malmendier, 2016).[21] The availability effect reflects the tendency to overweight information or experiences that are most easily recalled, particularly those that are vivid or emotionally charged. Each of the two dimensions described in Table 2 may affect availability in distinct ways. Second, our Appendix model implies a correlation between the two dimensions: accepters expect truths and rejecters expect lies. If the perception of a lie as a norm violation is associated with the expectation of truth, then the relevant gauge for a norm-violation-effect of a lie is for the accepters – for whom a lie is behavior contrary to what is expected for the majority of Senders.

Restricting attention to the accepters, by comparing LT-A to TT-A, also captures the understood content of a lie as described above, namely,

> (L1) "a false statement that deceives (LT-A) versus a true statement that is believed (TT-A)."

This content corresponds to related prior literature. For example, Brandts and Charness (2003) measure effects of deception (on punishment) by examining the effect of arriving at a disadvantageous payoff cell due to a believed-lie versus a believed-truth.[22] For our trust game, this comparison gives our main hypothesis:

[21] Also related are impulses to gravitate toward (and understate risks) of options that are more familiar (e.g., Schwartz and Song, 2009).

[22] Similarly, Sánchez-Pagés and Vorsatz (2007) identify effects of "lies that burn," versus "truths that don't burn," on punishment. A key difference is that these effects reflect payoff differences as well as lies vs. truths.

> *Hypothesis 2 (H2).* Hypothesis 1 holds for first-round accepters (LT-A versus TT-A).

Table 2 produces three other possible comparisons to judge effects of the *Lied To* versus *Told the Truth* experience. First is for first round rejecters, LT-R versus TT-R, comparing

> (L2) "a false statement that does not deceive (LT-R) versus a true statement that is not believed (TT-R)."

From the standpoint of the Receiver, (L2) does not align with the understood definition of a lie; it fails the "intention/deception" test requiring that the lie deceives. Moreover, because rejecters may expect lies (as in our Appendix model), the norm violation content of the treatment is in doubt and the anticipation effect – for those failing to anticipate the outcome correctly (the TT-R group) – may be more salient.[23] We therefore have no clear prediction for treatment effects on rejecters.

The second and third comparisons are between accepters and rejecters that control for anticipation outcomes, LT-R versus TT-A (the correct anticipators) or LT-A versus TT-R (the incorrect anticipators). Such comparisons are for different people (accepters vs. rejecters) for whom (i) social information can have different meaning and (ii) unobservables that drive first round behavior may be correlated with second round decisions. For example, our Appendix model predicts that rejecters have beliefs that Senders are predominantly untruthful; the social information indicating a majority (60 percent) of truthful Senders is therefore likely to be more salient. To address this potential confound, we can exploit the Control accepters and rejecters who are exposed to the same social information and whose first round behavior is driven by the same unobservables as are their LT and TT counterparts (due to random assignment to treatments). Treatment effects can then be inferred from difference in difference statistics, for example, the difference between LT-R and TT-A behavior less the corresponding difference between Control rejecter (C-R) and Control accepter (C-A) choices.

[23] An incorrect anticipation could negatively affect attitudes toward interactive play, and erode trust as a result. To the extent such effects are unmeasured, they could offset and crowd out the effect of the "lie experience" as we define it (in (L1)). To some extent, our results confirm this conjecture with our measured mood: The fraction of TT-R subjects experiencing post (vs. pre) treatment mood declines (26.3 percent) is significantly larger than for LT-R subjects (zero), with a z-statistic for the difference equal to 2.605 (p=0.015).

If incorrect anticipations sting, they could represent the more salient component of the treatment experience and confound effects of the norm violation versus compliance experience. Conversely, correct anticipations - precisely because treatment outcomes are as expected - are likely to be less salient relative to the norm violation content of the treatment experience. We therefore expect treatment effects consistent with Hypothesis 1 for the correct anticipators, but have no clear prediction for the incorrect anticipators.

> *Hypothesis 3 (H3).* For *correct anticipators* (LT-R and TT-A), the *Lied To* (vs. *Told the Truth*) treatment leads to a greater reduction in trust and trustworthiness when compared with corresponding differences in Control rejecter and Control accepter choices.

Hypothesis 3 is important as a check on Hypothesis 2. For the accepters, the LT treatment effect combines the two dimensions of Table 2, norm violation and incorrect anticipation, consistent with standard conceptions of lies. Hypothesis 3, if valid, identifies a pure norm violation effect of the treatments.

5 Implementation and Main Results

5.1 *Logistics*

The Receiver experiment is conducted in six one-shot sessions in Economics and Political Science classes at the University of California, Merced.[24] All subject responses are completely anonymous, with student participants identified for payment by registration numbers. No communication is allowed during the experiment. All three treatments are conducted in all classes, resulting in a sample of 266 subjects.[25] 81 subjects are exposed to the Control treatment; 91 subjects are exposed to the *Lied To* treatment; and 94 subjects are exposed

[24] Three Economics sessions are upper division and two are advanced lower division. The one Political Science session is lower division. We conduct the Sender side of the experiment in large introductory (prerequisite) Economics and business classes at U.C. Merced that minimize overlap with the Receiver sessions. Using course rosters, we ensure that no subject participates twice. Subjects are paid one week after the experiment is completed; to obtain payment, each student produces a tag with the registration number that is attached to the original questionnaire.

[25] The sample of 266 excludes two (LT) subjects who failed to complete the deception portion of the experiment.

to the *Told the Truth* treatment. Student payments, including a \$5 show-up fee (announced to subjects at the end of the experiment), average slightly above \$10.

The sample passes available balance checks across treatments, using the observable gender, initial mood, and accept decisions from the deception game. None of these variables is significantly different across treatments.[26] In addition (by design), the three treatments are equally represented in each of the classroom sessions, with sample distributions across the six courses that are statistically indistinguishable.[27]

5.2 *Main Results*

Table 3 and Figure 3 present proportions of subjects making trusting (Send) and trustworthy (Return \$7) decisions, in total and by treatment group (Control, Lied To, and Told the Truth), both for the full sample and for first round accepters and rejecters, respectively. Table 4 provides test statistics for Lied To (LT) versus Told the Truth (TT) treatment effects, three of which correspond with our Hypotheses (H1-H3).

Two conclusions are evident. First, *lies (vs. truths) significantly erode trust.* Overall, 31.9 percent of LT subjects choose Send, compared with 48.9 percent of TT subjects. Among accepters, the LT versus TT treatment lowers Send rates by 18.1 percentage points, slightly more than in the full sample. Both differences are statistically significant (Table 4), with p=0.0184 for the full sample difference and p=0.0306 for the accepters.[28] The raw data thus

[26] 49 percent of the overall sample is Male, with this proportion varying from 45.7 percent in the Control treatment to 48.9 and 52.7 percent in the LT and TT treatments, respectively. Initial mood averages 3.03 (where 3 is "good"), with tiny variation across treatments. "Accept" decisions are made by 74.4 percent of the sample, ranging from 71.6 percent in the Control to 71.4 and 79.8 percent in the LT and TT treatments. Differences across treatments are statistically insignificant and the acceptance rates themselves are in line with prior experiments. For example, 72 percent of Receivers accept in the Sutter (2009) experiments, 78 percent in Gneezy (2005), and 73.4 percent in Innes and Mitra (2013).

[27] We can construct three pairwise Kolmogorov-Smirnoff (KS) statistics for the null of a common course distribution across treatments. In each case, we order the classrooms so as to maximize the KS statistic. p-values for the three statistics are 0.96, .99 and .92, indicating common distributions.

[28] For (0,1) Bernoulli variables, test statistics for differences in means across samples/treatments are constructed as follows:

$$t = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{[s_1^2/N_1] + [s_2^2/N_2]}}$$

where $\overline{p}_i$ = sample i mean/proportion, $s_i^2 = \overline{p}_i(1-\overline{p}_i)$ = estimated variance for sample i draws. and N_i = sample i number of observations. Similarly, test statistics for difference-in-differences (sample 1 minus 2 support Hypotheses 1 and 2. For the "anticipators" - those who correctly anticipate a lie (LT-R) or truth (TT-A) - the Lied To (vs. Told the Truth) experience lowers the Send rate by 34.1 percent after netting out the predicted (Control) difference between rejecters and accepters. Again, the difference-in-difference is statistically significant (Table 4, p=0.0432), supporting Hypothesis 3.

Second, *lies (vs. truths) significantly erode trustworthiness.* In the full sample, 38.5 percent of LT subjects are trustworthy (choosing Return $7), compared with 60.6 percent of TT subjects, a difference of over 22 percentage points. For the accepters, the difference is even greater; 36.9 percent of LT-A subjects are trustworthy, compared with 64 percent of TT-A subjects, a difference of over 27 percentage points. From Table 4, both differences are statistically significant, with p=0.0026 in the full sample and p=0.0014 for the accepters. The difference-in-difference LT effect for the correct anticipators is also large (34.2 percent) and significant (p=0.05). These results support our three Hypotheses (H1-H3) with respect to subjects' Return decisions.

Perhaps a surprising feature of the results is the lower Send and Return rates for Control accepters versus rejecters. Although the differences are not significant (in part, we suspect, due to the limited number of rejecters in the sample), a possible explanation for this direction of effect is that the social information is more salient in promoting trust for rejecters. Rejecters don't "trust" in the Gneezy (2005) deception game, perhaps due to prior beliefs that Senders are unlikely to be "trustworthy". Learning that a majority (60 percent) of Senders are truthful can therefore motivate more trusting behavior. In contrast, first round "trusters" (the accepters) may find the social information uninformative or indicative of less

versus sample 3 minus 4) are constructed as

$$t = \frac{(\overline{p}_1 - \overline{p}_2) - (\overline{p}_3 - \overline{p}_4)}{\sqrt{\sum_{i=1}^{4}[s_i^2/N_i]}}$$

For example, to construct the difference-in-difference statistic for the correct anticipators, sample 1 is the LT-R, 2 is the TT-A, 3 is the C-R, and 4 is the C-A. For continuous variables (such as beliefs), t statistics are calculated in the same way, with sample means and variances taking the place of $\overline{p}_i$ and s_i^2. By the Central Limit Theorem and the Law of Large Numbers, these statistics are approximately distributed standard Normal under the nulls of zero difference / difference-in-difference. However, following convention, we construct p-values using the t distribution with a conservative degrees of freedom count set equal to the minimum sample size (N_i) minus one.

truthfulness than expected. [29]

We note that our TT subjects are much more trusting and trustworthy in the trust game relative to the Controls, whereas the LT subjects are only slightly less trusting and trustworthy compared with the Controls (particularly for the accepters). In this sense, it seems that truth promotes trust more than lies deter it. However, such comparisons may have more to do with how social information (on propensities for honesty) affect the Controls and less to do with the LT vs. TT treatment effects. For this reason, we focus on direct comparisons between LT and TT subject decisions.

6 Robustness, the Interpretation of Lies, and Beliefs

6.1 *Regressions*

Regressions enable more precision in the estimation of treatment effects, with controls for course effects, gender, and of particular importance, mood. Table 5 describes the proportions of the sample, overall and by treatment, that report a positive or negative mood change post-treatment (vs. pre-treatment). Over 25 percent of treated (LT and TT) subjects report an improvement or decrease in mood. Lied To (LT) subjects exhibit a lower propensity for positive mood change, and a higher propensity for negative mood change, when compared with their Told the Truth (TT) counterparts. As a result, there is a significant difference between LT and TT subjects in the extent to which the rate of negative mood change exceeds the rate of positive mood change (last column, Table 5, p=0.0128).[30] As indicated in the Appendix, this tendency is particularly acute for first round accepters, for whom LT (vs. TT) subjects experienced a 17.5 percent greater frequency of negative mood change and

[29] Unfortunately, a direct test of this conjecture – by elicitation and study of beliefs for the Gneezy (2005) game – is not possible in the experiment. We eschewed pre-treatment belief elicitation due to the potential for contaminating treatment effects. And post-treatment elicitation is precluded by the social information that we provide in the treatments.

[30] The reported effect in the last column of Table 5 is the difference-in-difference,

$$(\overline{p}_{LT,+} - \overline{p}_{LT,-}) - (\overline{p}_{TT,+} - \overline{p}_{TT,-})$$

where $\overline{p}_{t,c}$ = proportion of subjects reporting positive ($c = +$) or negative ($c = -$) mood change in the $t = LT$ or $t = TT$ sample.

an 18.4 percent lower frequency of positive mood change. Are treatment effects on trust decisions attributable to these changes in mood?

Table 6 reports regressions to explain subjects' Send and Return decisions, controlling for course effects, gender, initial mood, and mood change. Estimated models have the form:

$$y_i = \alpha + \beta' T_i + \eta' X_i + \epsilon_i \tag{1}$$

where y_i = (0,1) outcome for subject i (Send or Return \$7), T_i = vector of (0,1) treatments for observation i, X_i = vector of controls, and ϵ_i = random disturbance.[31] For the difference (LT vs. TT) models in the top three panels of Table 6, T_i includes the Control indicator and the Lied To indicator (with TT as baseline). The reported Lied To effect corresponds with the estimated marginal effect of the LT indicator.[32] In the difference-in-difference models, T_i includes the five accept and reject specific treatments (LT-A, LT-R, TT-A, TT-R, and Control-R), excluding the baseline Control-A. The corresponding Lied To effect for the correct anticipators is given by the difference in estimated marginal effects, LT-A - TT-R - C-R, and the p-value is for an F or χ^2 test of a zero difference-in-difference.[33] More controls are added when moving from left to right in the table, starting with course effects and gender, adding initial mood indicators, and finally adding accept and reject specific indicators for positive and negative mood change.[34]

Estimated magnitudes and significance of treatment effects are similar across the models and robust to all controls. The only mood variable that is statistically significant in the full sample regressions is negative mood change for the rejecters (which reduces trust and trustworthiness).[35] For the accepters, none of the mood variables is significant (with lowest p-value equal to 0.295). Overall and for accepters, we estimate that being Lied To reduces rates of trust by 17 to 19 percentage points and rates of trustworthiness by 23 to 31 percentage

[31] In the probit models, y_i represents a continuous index value, rather than the (0,1) outcome variable. In all models, we provide consistent estimates of standard errors by adjusting for generalized heteroskedasticity.

[32] Corresponding p-values are determined using the t(df=N-K) distribution, where K=number of regressors.

[33] An F test is presented for the linear models and a χ^2 test for the Probit.

[34] The on-line Appendix provides full regression results for models reported in Tables 6 to 8.

[35] Corresponding p-values are 0.083/0.157 in the Send models (columns (3)/(4) in Table 6), and 0.0012/0.0172 in the Return models ((3)/(4)).

points – estimates we interpret as experience effects of receiving a lie (vs. a truth). With the full complement of controls, there is also a large negative estimated effect of the Lied To (vs. Told the Truth) treatment on trust decisions of rejecters, albeit statistically insignificant (with p>0.19). The results indicate that the deleterious effects of the Lied To experience on trust are robust in general and, in particular, not attributable to treatment-induced changes in mood.

6.2 *The Interpretation of Lies*

The design intent of our experiment is to present subjects with the experience of a norm violation (a lie) versus norm compliance (a truth). To determine whether we succeed in doing so, we ask subjects (at the very end of the experiment) to indicate the normative content of a false message in the deception game from their perspective. Table 7 presents percentages of subjects indicating that sending a false message represents a norm violation in each of the following statements: disagreeing that it is "not really lying, just being rational" (S1), agreeing that it is "trying to trick/deceive the Receiver" (S2), or agreeing that it is "not the right thing to do". We also ask subjects (in Q4) to predict the proportion of participants agreeing with the normative statement S3 (paying $1 for a correct prediction, within a 15 percentage point band).

Subjects' answers reveal that substantial majorities find normative content in a false message. Indeed, over 94 percent of the full sample indicate that a false message represents a norm violation in at least one of the three statements S1-S3, and over 76 percent do so in two or all of the statements. The majorities are large, significant ($p < 0.01$), and do not exhibit significant variation across treatments (see Appendix). The mean belief on the norm that a false message is "not the right thing to do" (Q4) is over two-thirds.

Because our hypothesized treatment effects H1-H3 are driven by the premise that a lie (as defined) has normative content, we also estimate the LT (vs. TT) effect on trust choices for samples restricted to subjects indicating so in their answers to S1, S2, S3, and Q4 (with beliefs greater than 55 percent), respectively. The bottom panel of Table 7 presents the corresponding estimates (using the full set of controls). The results again support our hypotheses, with magnitudes of effect that are similar to those from the full sample (c.f., Table 6, model 3).

6.3 *The Role of Beliefs*

Sapienza, Toldra-Simats and Zingales (STZ, 2013) suggest that the best measure of trust that one can obtain from the Berg, et al. (1995) game is one based on expectations: For a given amount of money sent to a Returner, how much money does a Sender expect to be returned? For our simplified trust game, this question is addressed with a measure of how likely a subject believes it is that a Returner will choose the generous return strategy. As described in Section 3, we use an incentive compatible approach to elicit each subject's beliefs about the fraction of participants who Send (Q1) and the fraction who Return $7 (Q2). Q2 gives the STZ measure of trust for our game.

Table 8 presents mean predictions and estimated (LT vs. TT) treatment effects for Q1 and Q2. Mean beliefs are 43.3 percent for Send and 47.6 percent for Return $7, strikingly close to actual outcomes in the experiment, 39.5 percent for Send and 47.7 percent for Return $7 (Table 3). Perhaps more important, the "Lied To" treatment has a negative effect on beliefs about both trust (Send) and trustworthiness (Return). For the full sample, the LT effect is statistically significant for both Q1 and Q2 (column 2 of Table 8, p=0.017 for Send and p=0.059 for Return). For the accepters, the effect on Q2 is significant (column 3, p=0.071), while for the correct anticipators, the (difference-in-difference) effect is only significant for Q1 (column 4, p=0.005). Overall, Table 8 indicates that the experience of a lie (vs. a truth) erodes the STZ measure of trust.[36] These results might reflect false consensus effects – subject beliefs that conform to the subjects' own choices (Ross, et al., 1977; Ellingsen, et al., 2010; Butler et al., 2015).

[36] We cannot rule out some attenuation of treatment effects on beliefs due to strategic hedging (Blanco et al., 2010). For example, those who Send might predict a lower fraction of generous returners in order to hedge the risk of facing a stingy returner in the Sender role. However, any such effects would reflect very weak beliefs and bias against the treatment effects that we find. (Hedging effects could only go in one direction in our experiment, affecting beliefs and not trust decisions, because the belief elicitation is not known to the subjects until the end of the experiment.)

7 Conclusion

We find that being on the receiving end of a lie (vs. a truth) leads to an erosion of trust, even in interactions with those who have nothing to do with the initial deception and even though the deceptive act is known to have no bearing on the overall propensity for dishonesty among experimental participants. Given the central role that trust is known to play in promoting economic interchange and growth, this conclusion suggests that social institutions that deter dishonesty and promote norms of truthfulness are of potential economic value.

Our results are likely to understate the impact of lies in real life. In our experiment, each Message is from an anonymous Sender who has made a decision in a different class at a different time, having no knowledge of who would receive the communication. Less anonymity and distance, we believe, could be expected to elevate the sting from a lie.

A key feature of the analysis is the identification of an individual experience effect of the Lied To and Told the Truth treatments, controlling for potential treatment effects on mood, payoffs, and overall Sender propensities for honesty. Separate from everything else, the individual experience of a norm violation (a lie) alters behavior. These results expose a potentially general link between individual experience and behavior in social interchange. A great deal of research studies what drives or deters trust and trustworthiness, including (among others) expectations (Sapienza, et al., 2013), reciprocity (Charness and Rabin, 2002), and guilt aversion (Charness and Dufwenberg, 2006). One possible interpretation of our results is that reciprocal preferences that drive trust are determined by a broad social context and specific experiences in a compendium of social interactions, including experiences of norm violations such as lies; lies may reduce reciprocation and/or the extent of guilt aversion that lead to trustworthy decisions.

8 References

Abeler, J., Becker, A. and Falk, A. (2014). "Representative evidence on lying costs." *Journal of Public Economics*, 113, 96-104.

Alexander, R. (1987). *The Biology of Moral Systems.* Aldine-de-Gruyter: New York.

Al-Ubaydli, O., Houser, D., Nye, J., Paganelli, M. and Pan, X. (2013). "The causal effect of market priming on trust." *PLOS ONE*, 8(3), 1-8.

Altmann, S., Dohmen, T. and Wibral, M. (2008). "Do the reciprocal trust less?" *Economics Letters*, 99(3), 454-457.

Andreoni, J. (1990). "Impure altruism and donations to public goods: A theory of warm-glow giving." *Economic Journal*, 100(401), 464-477.

Andreoni, J. and Bernheim, D. (2009). "Social image and the 50-50 norm." *Econometrica*, 77(5), 1607-1636.

Ashraf, N., Bohnet, I. and Piankov, N. (2006). "Decomposing trust and trustworthiness." *Experimental Economics*, 9(3), 193-208.

Bartling, B., Engl, F. and Weber, R. (2014). "Does willful ignorance deflect punishment? An experimental study." *European Economic Review*, 70, 512-524.

Bartling, B. and Fischbacher, U. (2012). "Shifting the blame: On delegation and responsibility." *The Review of Economic Studies*, 79 (1), 67-87.

Battigalli, P., Charness, G. and Dufwenberg, M. (2013). "Deception: The role of guilt." *Journal of Economic Behavior and Organization*, 93, 227-233.

Battigalli, P. and Dufwenberg, M. (2007). "Guilt in games." *American Economic Review*, 97, 170-176.

Battigalli, P. and Dufwenberg, M. (2009). "Dynamic psychological games." Journal of Economic Theory, 97, 1-35.

Bellemare, C. and Kroger, S. (2007). “On representative social capital.” *European Economic Review*, 51(1), 183-202.

Ben-Ner, A., Putterman, L., Kong, F., and Magan, D. (2004). “Reciprocity in a two-part dictator game.” *Journal of Economic Behavior and Organization*, 53, 333-352.

Berg, J., Dickhaut, J., and McCabe, K. (1995). “Trust, reciprocity, and social-history.” *Games and Economic Behavior*, 10(1), 122-142.

Blanco, M., Engelmann, D., Koch, A., and Normann, H. (2010). “Belief elicitation in experiments: Is there a hedging problem?” *Experimental Economics*, 13, 412-38.

Bok, S. (1978). *Lying: Moral Choice in Public and Private Life.* New York.Pantheon Books.

Brandts, J. and Charness, G. (2011). “The strategy versus the direct-response method: A first survey of experimental comparisons.” *Experimental Economics*, 14(3), 375-398.

Brandts, J. and Charness, G. (2003). “Truth or consequences: An experiment.” *Management Science*, 49(1): 116-130.

Burks, S., Carpenter, J. and Verhoogen, E. (2003). “Playing both roles in the trust game.” *Journal of Economic Behavior and Organization*, 51, 195-216.

Butler, J., Giuliano, P. and Guiso, L. (2015). “Trust, values and false consensus.” *International Economic Review*, 56(3), 889-915.

Cai, J. and Song, C. (2011). “Insurance take-up in rural China: learning from hypothetical experience.” Working Paper, U.C. Berkeley.

Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton Univ. Press: Princeton, New Jersey.

Camerer, C. and Ho, T. (1999). “Experience-weighted attraction learning in normal form games.” *Econometrica*, 67, 827-74.

Capra, C. M. (2004). "Mood-driven behavior in strategic interactions." *AEA Papers and Proceedings, American Economic Review*, 95(2), 367-372.

Charness, G. and Dufwenberg, M. (2006). "Promises and partnership." *Econometrica*, 74(6), 1579-1601.

Charness, G. and Dufwenberg, M. (2010). "Bare promises: An experiment." *Economics Letters*, 107, 281-283.

Charness, G. and M. Rabin (2002). "Understanding social preferences with simple tests." *Quarterly Journal of Economics*, 117, 817-868.

Charness, G. and M. Rabin (2005). "Expressed preferences and behavior in experimental Games." *Games and Economic Behavior*, 53, 151-169.

Charness, G. and Sutter, M. (2012). "Groups make better self-interested decisions." *Journal of Economic Perspectives*, 26(3), 157-176.

Chaudhuri, A. and Gangadharan, L. (2007). "An experimental analysis of trust and trustworthiness." *Southern Economic Journal*, 73(4), 959-985.

Costa-Gomez, M., Huck, S. and Weizsacker, G. (2010). "Beliefs and actions in the trust game." *IZA Disc. Paper* 4709.

Cox, J., Friedman, D. and Sadiraj, V. (2008). "Revealed altruism." *Econometrica*, 76(1), 31-69.

Dana, J., Weber, R. and Kuang, J. (2007). "Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness." *Economic Theory*, 33(1), 67-80.

DePaulo, B., D. Kashy, S. Kirkendol, N, Wyer, M. and Epstein, J. (1996). "Lying in everyday life." *Journal of Personality and Social Psychology*, 70(5), 979-995.

Dohmen, T., Falk, A., Huffman, D. and Sunde, U. (2012). "The intergenerational transmission of risk nad trust attitudes." *Review of Economic Studies*, 79(2), 645-677.

Dreber, A. and Johannesson, M. (2008). "Gender differences in deception." *Economics Letters*, 99, 197-199.

Duffy, J. and Feltovich, N. (2006). "Words, deeds, and lies: Strategic behavior in games with multiple signals." *Review of Economic Studies*, 73, 669-688.

Dufwenberg, M. and Gneezy, U., (2000). "Measuring beliefs in an experimental lost wallet game." *Games and Economic Behavior*, 30, 163-182.

Dufwenberg, M., Gneezy, U., Guth, W., and van Damme, E. (2001). "Direct vs indirect reciprocity: An experiment." *Homo Oeconomicus*, 18, 19-30.

Dufwenberg, M. and Kirchsteiger, G. (2004). "A theory of sequential reciprocity." *Games and Economic Behavior*, 47, 268-298.

Ellingsen, T., Johannesson, M., Lilja, J. and Zetterqvist, H. (2009). "Trust and truth." *Economic Journal*, 119, 252-276.

Ellingsen, T., Johannsesson, M., Tjotta, S., and Torsvig, G. (2010). "Testing guilt aversion." *Games and Economic Behavior*, 68, 95-107.

Erat, S. and Gneezy, U. (2012). "White lies." *Management Science*, 58 (4), 723-733.

Fehr, E., Fischbacher, U., von Rosenbladt, B., Schupp, J., and Wagner, G. (2003). "Nationwide laboratory examining trust and trustworthiness by integrating behavioural experiments into representative surveys." *CEPR Disc. Paper* 3858.

Fehr, E. and Schmidt, K.M. (1999). "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114, 817-868.

Fischbacher, U. and Fllmi-Heusi, F. (2013). "Lies in disguise An experimental study on cheating." *Journal of the European Economic Association*, 11(3), 525-547.

Freeman, R. and Gelber, A. (2010). "Prize structure and information in tournaments: Experimental evidence." *American Economic Journal: Applied Econ.*, 2(1), 149-164.

Gibson, R., Tanner, C., and Wagner, A. F. (2013). "Preferences for truthfulness: Heterogeneity among and within individuals." *American Economic Review*, 103(1), 532-548.

Glaeser, E., Laibson, D., Scheinkman, J. and Soutter, C. (2000). "Measuring trust." *Quarterly Journal of Economics*, 65(3), pp. 811-846.

Gneezy, U. (2005). "Deception: The role of consequences." *American Economic Review*, 95, 384-394.

Gneezy, U., Rockenbach, B. and Serra-Garcia, M. (2013). "Measuring lying aversion." *Journal of Economic Behavior and Organization*, 93, 293-300.

Greiner, B., and Levati, V.M. (2005). "Indirect reciprocity in cyclical networks: An experimental study." *Journal of Economic Psychology*, 26, 711-731.

Grossman, Z. (2015). "Strategic ignorance and the robustness of social preferences." *Management Science*, 60(11), 2659-2665.

Guiso, L., Sapienza, P. and Zingales, L. (2004). "The role of social capital in financial development." *American Economic Review*, 94, 526-556.

Guiso, L., Sapienza, P. and Zingales, L. (2008a). "Trusting the stock market." *Journal of Finance*, 63, 2557-2600.

Guiso, L., Sapienza, P. and Zingales, L. (2008b). "Social capital as good culture." *J. of European Econ. Assoc.*, 6, 295-320.

Guiso, L., Sapienza, P. and Zingales, L. (2009). "Cultural biases in economic exchange." *Quarterly Journal of Economics*, 124(3), 1095-1131.

Herne, K., Lappalainen, O. and Kestil-Kekkonen, E. (2013). "Experimental comparison of direct, general, and indirect reciprocity." *Journal of Socio-Economics*, 45, 38-46.

Hertwig, R., Barron, G. and Weber, E. (2004). "Decisions from experience and the effect of rare events in risky choice." *Psychological Science*, 15, 534-39.

Houser, D., Schunk, D. and Winter, J. (2010). “Distinguishing trust from risk: an anatomy of the investment game.” *J. of Econ. Behavior and Organization*, 74(1–2), 72-81.

Houser, D., Vetter, S. and Winter, J. (2012). “Fairness and cheating.” *European Economic Review*, 56, 1645-1655.

Hurkens, S. and Kartik, N. (2009). “Would I lie to you? On social preferences and lying aversion.” *Experimental Economics*, 12(2), 180-192.

Innes, R. and Mitra, A. (2013). “Is dishonesty contagious?” *Economic Inquiry*, 51(1), 722-734.

Johnson, N.D. and Mislin, A.A. (2011). “Trust games: A meta-analysis.” *Journal of Economic Psychology*, 32, 865-889.

Josephson Institute of Ethics (2012). *2012 Report Card on the Ethics of American Youth.* Los Angeles: Josephson Institute.

Keizer, K., Lindenberg, S. and Steg, L. (2008). “The spreading of disorder.” *Science*, 322, 1681-1685.

Kirchsteiger, G., Rigotti, L. and Rustichini, A. (2006). “Your morals might be your moods.” *Journal of Economic Behavior and Organization*, 59, 155-172.

Knack, S. and Keefer, P. (1997). “Does social capital have an economic payoff? A cross-country investigation.” *Quarterly Journal of Economics*, 112, 1252-1288.

La Porta, R., Lopez de Silanes, F., Shleifer, A. and Vishny, R. (1997). “Trust in large organisations.” *American Economic Review*, 87(2), 333-338.

Lazzarini, S., Madalozzo, R., Artes, R. and de Oliveira Siqueira, J. (2004) “Measuring trust: An experiment in Brazil.” *Ibmec working paper WPE*, 2004.

Levine, D. (1998). “Modeling altruism and spitefulness in experiments.” *Review of Economic Dynamics*, 1(3), 593-622.

Lundquist, T., Ellingsen, T., Gribbe, E., and Johannesson, M. (2009). "The aversion to lying." *Journal of Economic Behavior and Organization*, 70(1-2), 81-92.

Mahon, J. (2016). "The definition of lying and deception." *The Stanford Encyclopedia of Philosophy*, Spring 2016 Edition, Edward N. Zalta, ed..

Malmendier, U. (2016). "Experience effects." Keynote Address, European ESA Meetings, Heidelberg.

Malmendier, U. and Nagel, S. (2011). "Depression babies: Do macreconomic experiences affect risk taking?" *Quarterly Journal of Economics*, 126, 373-416.

Malmendier, U. and Nagel, S. (2016). "Learning from inflation experiences." *Quarterly Journal of Economics*, in press.

Malmendier, U., Tate, G., and Yan, J. (2011). "Overconfidence and early-life experiences: The effect of managerial traits on corporate financial policies." *Journal of Finance*, 66, 1687-1733.

Mazar, N., Amir, O., and Ariely, D. (2008). "The dishonesty of honest people: A theory of self-concept maintenance." *Journal of Marketing Research*, 45, 633-644.

Nesbit, R. and Ross, L. (1980). *Human Inference.* Prentice-Hall: Englewood Cliffs, New Jersey.

Rabin, M. (1993). "Incorporating fairness into game theory and economics." *American Economic Review*, 83, 1281-1302.

Rode, J. (2010). "Truth and trust in communication: Experiments on the effect of a competitive context." *Games and Economic Behavior*, 68, 325-338.

Rosenbaum, S., Billinger, S. and Stieglitz, N. (2014). "Lets be honest: A review of experimental evidence of honesty and truth-telling." *J. of Econ. Psychology*, 45, 181-196.

Ross, L., Greene, D., and House, P. (1977). "The false consensus effect: An egocentric bias in social perception and attribution processes." *J. of Exper. Soc. Psych.*, 13, 279-301.

Sánchez-Pagés, S. and Vorsatz, M. (2007). “An experimental study of truth-telling in a senderreceiver game.” *Games and Economic Behavior*, 61, 86-112.

Sánchez-Pagés, S. and Vorsatz, M. (2009). “Enjoy the silence: An experiment on truth-telling.” *Experimental Economics*, 12 (2), 220-241.

Sapienza, P., Toldra-Simats, A. and Zingales, L. (2013). “Understanding trust.” *Economic Journal*, 123, 1313-1332.

Schweitzer, M., Hershey, J., and Bradlow, E. (2006). “Promises and lies: Restoring violated trust.” *Organizational Behavior and Human Decision Processes*, 101, 1-19.

Simonsohn, U., Karlsson, N., Loewenstein, G., and Ariely, D. (2008). “The tree of experience in the forest of information: Overweighing experienced relative to observed information.” *Games and Economic Behavior*, 62, 263-86.

Song, H. and Schwartz, N. (2009). “If it‘s difficult to pronounce, it must be risky.” *Psychological Science*, 20, 135-8.

Stanca, L. (2009). “Measuring indirect reciprocity: Whose back do we scratch?” *Journal of Economic Psychology*, 30, 190-202.

Sutter, M. (2009). “Deception through telling the truth? Experimental evidence from individuals and teams.” *Economic Journal*, 119, 47-60.

Tversky, A. and Kahneman, D. (1974). “Judgment under uncertainty: Heuristics and biases.” *Science*, 185, 1124-31.

Tyler, J.M., Feldman, R. S., and Reichert, A. (2006). “The price of deceptive behavior: Disliking and lying to people who lie to us.” *Journal of Experimental and Social Psychology*, 42, 69-77.

Zak, P. and Knack, S. (2001). “Trust and growth.” *Economic Journal*, 111, 295-321.

Appendix: Toy Model of Gneezy (2005) Deception Game

Consider the Gneezy (2005) game in Section 3.1. Assume

(1) Two types of risk neutral Senders: zero lie averse (type O) and sufficiently lie averse that truth-telling is desired (type A).

(2) Two types of risk neutral Receivers: High belief (H) and low belief (L). H (L) types believe the proportion of lie averse Senders is $p_H > 0.5$ ($p_L < 0.5$). The proportion of H types is $g > 0.5$.

(3) The types of Senders, types of Receivers, and g are public information.

(4) Receivers know that Senders have conflicting incentives. That is, excluding lie aversion, the Sender's monetary-equivalent benefit of the "lie option" (vs. "truthful option") is $b > 0$. The monetary-equivalent benefit of the "truthful" (vs. "untruthful") option to the Receiver is $c > 0$. The Receivers know that $b > 0$, $c > 0$, and $b < a =$ monetary-equivalent lie aversion for A Senders; they need not know the specific values of a, b, or c.

For pure strategies, let $r_i \in \{0, 1\}$ be the "accept" ($r = 1$) or "reject" ($r = 0$) decision of a Receiver of type $i \in \{H, L\}$, and $s_j \in \{0, 1\}$ be the "truthful" ($s = 1$) or "untruthful" ($s = 0$) message choice of a Sender of type $j \in \{O, A\}$. Net expected payoffs are:

$$\text{Payoff to Sender O} = \pi^S_{Ol} = [gr_H + (1-g)r_L]b \qquad \text{under "lie"} \qquad \text{(A1)}$$

$$= \pi^S_{Ot} = [g(1-r_H) + (1-g)(1-r_L)]b \qquad \text{under "truth"} \qquad \text{(A2)}$$

$$\text{Payoff to Receiver i} = \pi^R_{ia} = [p_i s_A + (1-p_i)s_O]c \qquad \text{under "accept"} \qquad \text{(A3)}$$

$$= \pi^R_{ir} = [p_i(1-s_A) + (1-p_i)(1-s_O)]c \qquad \text{under "reject"} \qquad \text{(A4)}$$

Observation 1. If $g > 0.5$, the unique Nash Equilibrium is: $s_A = 1$, $s_O = 0$, $r_H = 1$, $r_L = 0$.

Proof. By construction, $s_A = 1$. Therefore, (with $1 \geq p_H > 0.5$), $\pi^R_{Ha} > \pi^R_{Hr}$ and, hence, $r_H = 1$. With $r_H = 1$ and $g > 0.5$, $\pi^S_{Ol} > \pi^S_{Ot}$ and, hence, $s_O = 0$. Finally, with $p_L < 0.5$,

$s_O = 0$, and $s_A = 1$, $\pi^R_{La} < \pi^R_{Lr}$ and, hence, $r_L = 0$.

Discussion. With $g > 0.5$, the model produces a unique pure strategy Nash Equilibrium in which

(i) a majority of Receivers ($g > 0.5$) "accept,"

(ii) a minority of Receivers ($(1 - g) < 0.5$) "reject,"

(iii) some Senders (A types) are truthful,

(iv) some Senders (O types) lie,

(v) Receivers are distinguished by their beliefs about Sender truthfulness,

(vi) "accepting" (H type) Receivers are surprised by a lie (expecting a truth with probability $p_H > 0.5$),

(vii) "rejecting" (L type) Receivers are surprised by a truth (with $p_L < 0.5$) and by social information indicating a majority of truthful Senders.

The Case of $g < 0.5$. For mixed strategies, payoffs are as in (A1)-(A4) with r_i and s_O representing probabilities.

Observation 2. If $g < 0.5$, then the unique Nash Equilibrium is: $s_A = 1$, $r_H = 1$, $r_L = \left(\frac{1}{2}\right)\left(\frac{1-2g}{1-g}\right) \in (0, 0.5)$ and $s_O = \left(\frac{1}{2}\right)\left(\frac{1-2p_L}{1-p_L}\right) \in (0, 0.5)$.

Proof. (I) *There is no N.E. in pure strategies.* If $r_L = 0$ (and $g < 0.5$), $\pi^S_{Ol} < \pi^S_{Ot}$ and, hence, $s_O = 1$; however, with $s_O = s_A = 1$, $\pi^R_{La} > \pi^R_{Lr}$ and, hence, $r_L = 1$ (a contradiction). Suppose instead that $r_L = 1$ (and $g < 0.5$). Then $\pi^S_{Ol} > \pi^S_{Ot}$ and, hence, $s_O = 0$; however, with $s_O = 0$ and $s_A = 1$ (and $p_L < 0.5$), $\pi^R_{La} < \pi^R_{Lr}$ and, hence, $r_L = 0$ (a contradiction).

(II) *Observation 2.* By the proof of (I), r_L must be interior in any N.E., implying (with $s_A = 1$) that

$$\pi^R_{ia} - \pi^R_{ir} \,=\, 0 \;\leftrightarrow\; p_i + (1 - p_i)(2s_O - 1) \,=\, 0 \text{ for i=L} \tag{A5}$$

With $p_H > p_L$, equation (A5) implies (i) $\pi^R_{Ha} > \pi^R_{Hr}$ and (ii) $s_O = \left(\frac{1}{2}\right)\left(\frac{1-2p_L}{1-p_L}\right) \in (0, 0.5)$. By (i), $r_H = 1$, and by (ii), s_O is interior and must solve: $\pi^S_{Ol} - \pi^S_{Ot} = 0 \leftrightarrow g + (1-g)(2r_L - 1) = 0 \leftrightarrow r_L = \left(\frac{1}{2}\right)\left(\frac{1-2g}{1-g}\right) \in (0, 0.5)$ for $g < 0.5$.

Corollary. If $g < 0.5$ and the true fraction of lie averse Senders is p_T, then in the N.E., the overall probability of a Receiver accepting is one-half and the overall probability of a truthful message is:

$$p^* = p_T + (1 - p_T)\left(\frac{1 - 2p_L}{1 - p_L}\right)\left(\frac{1}{2}\right) > p_T \tag{A6}$$

Proof. The accept probability is $g + (1-g)r_L = 0.5$. Eq. (A6) follows from Obs. 2.

For example, if $g < 0.5$, $p_L = \frac{1}{3}$, and $p_T = 0.5$, the overall probability of truth is $p^* = \left(\frac{1}{4}\right) + \left(\frac{3}{4}\right) p_T = \frac{5}{8}$.

Figure 1. Timeline for Experiment

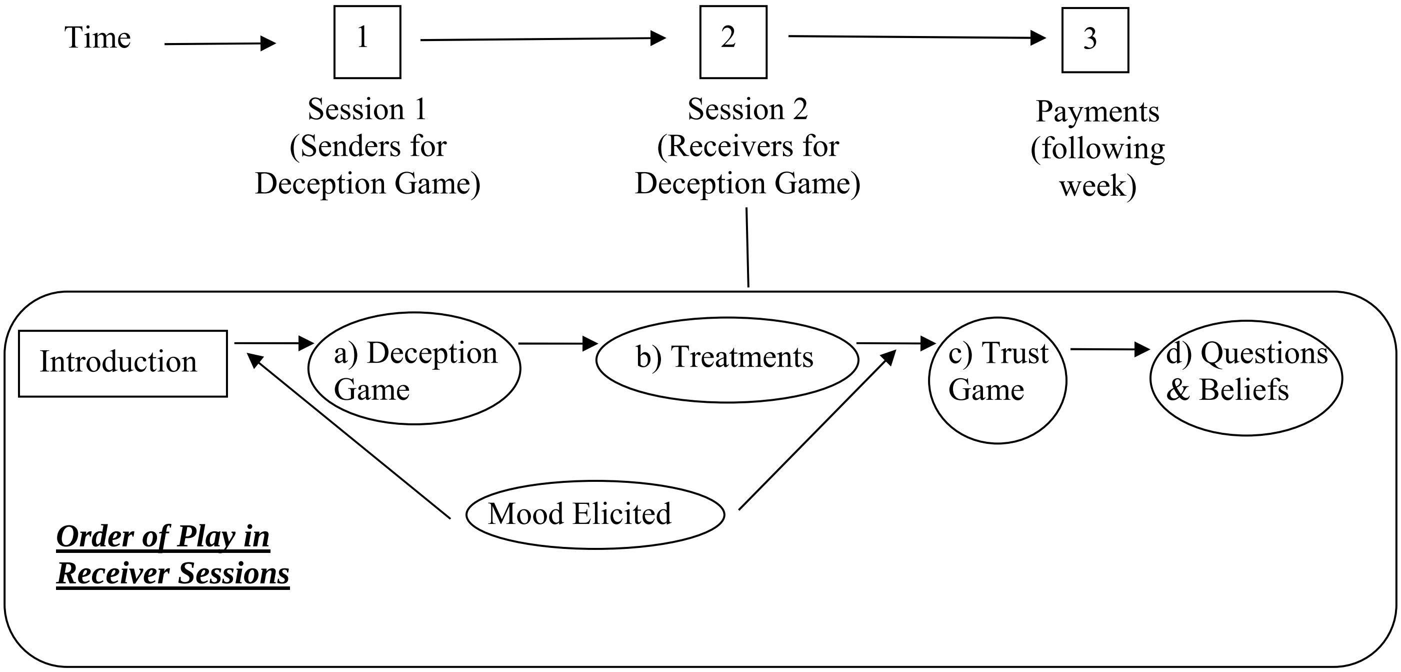


Figure 2A. Gneezy Game and Receiver Treatment

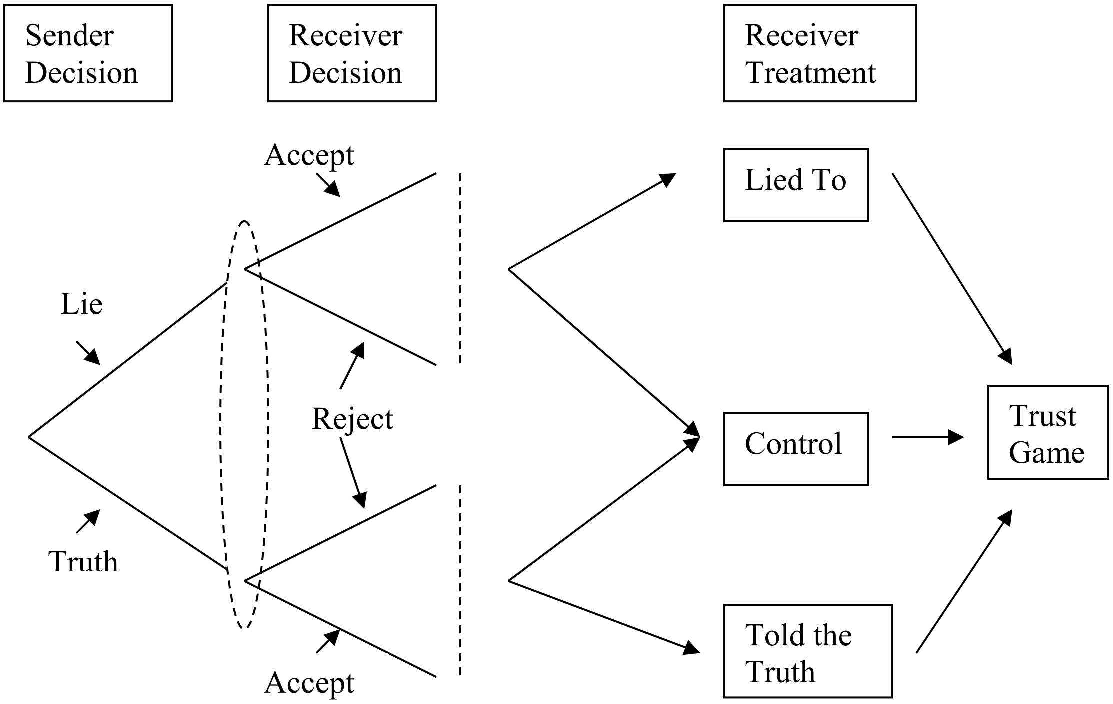
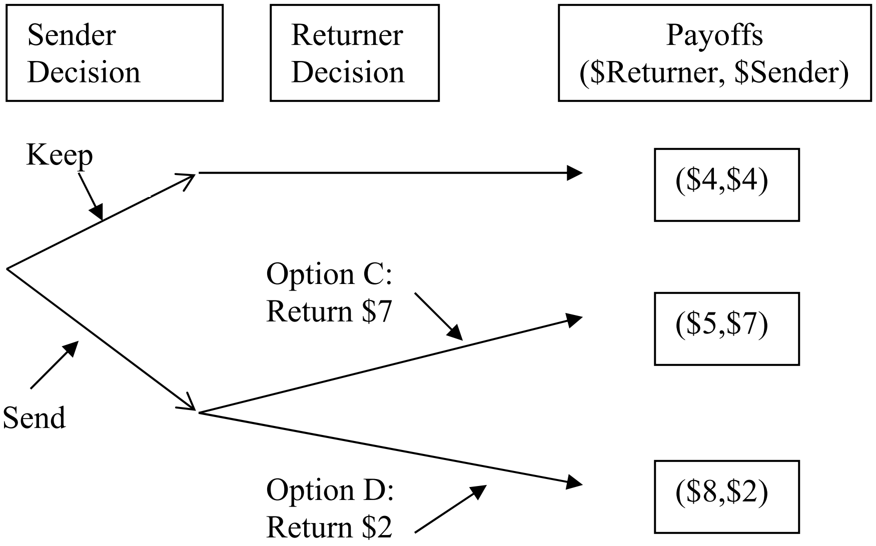


Figure 2B. Trust Game, Gneezy Receivers Play Both Roles



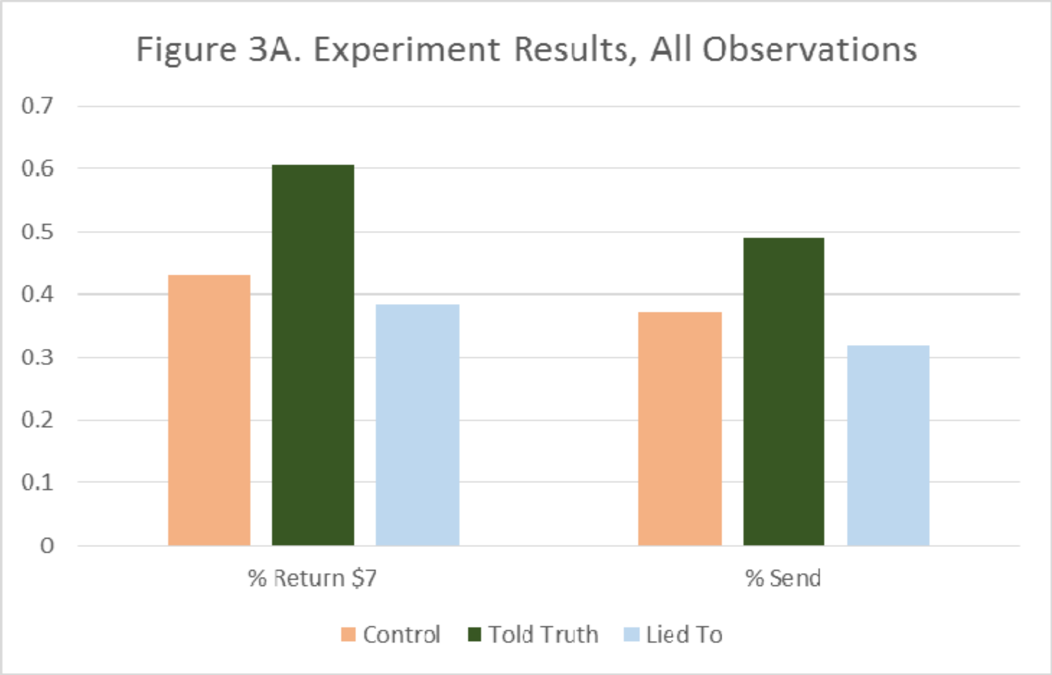


Figure 3A. Experiment Results, All Observations

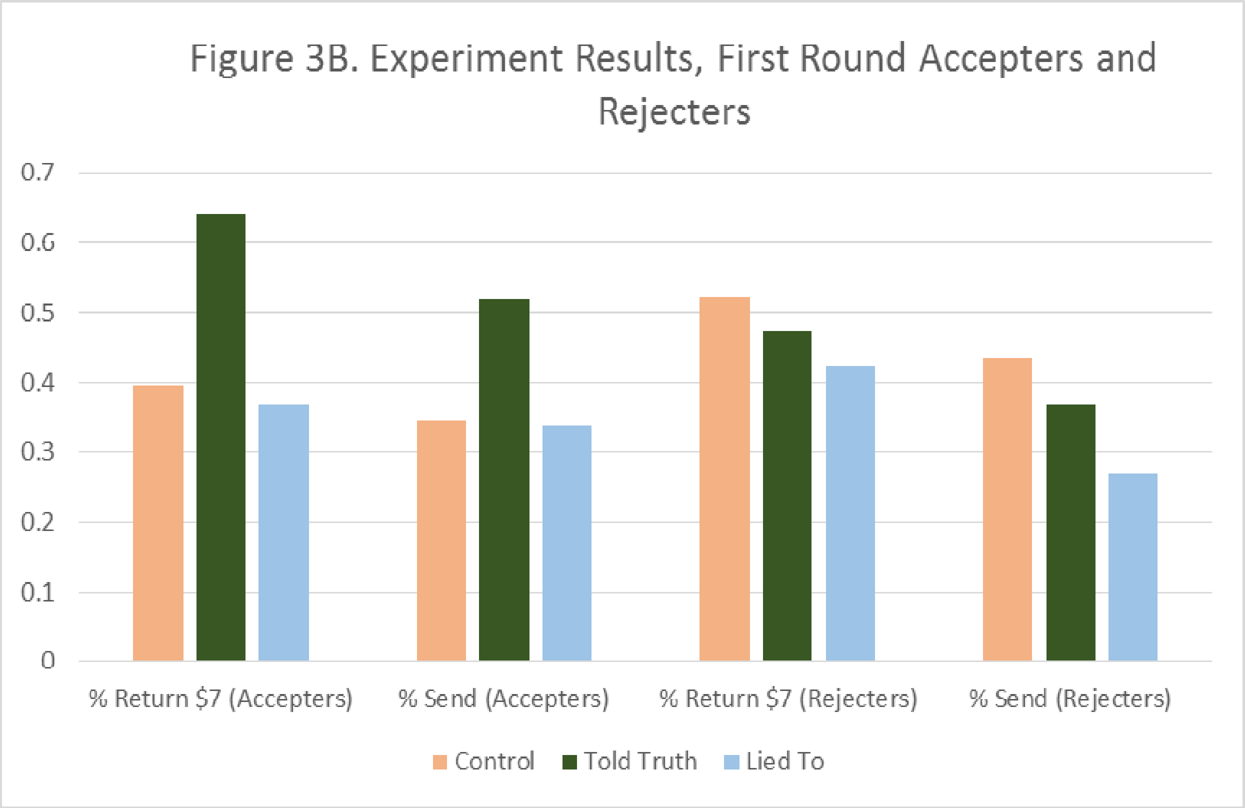


Figure 3B. Experiment Results, First Round Accepters and Rejecters

Table 3. Sample Summary Statistics for the Trust Game

	All Observations	Treatments		
		Control	Lied To	Told the Truth
Full Sample, N →	266	81	91	94
Send / Trust	0.395	0.370	0.319	0.489
Return $7 / Trustworthy	0.477	0.432	0.385	0.606
First Round Accepters, N →	198	58	65	75
Send / Trust	0.409	0.345	0.338	0.520
Return $7 / Trustworthy	0.480	0.397	0.369	0.640
First Round Rejecters, N →	68	23	26	19
Send / Trust	0.353	0.435	0.269	0.368
Return $7 / Trustworthy	0.471	0.522	0.423	0.474

N=number of observations. Summary statistics represent proportions of each sample making the indicated choice.

Table 4. Difference Statistics: "Lied To" vs. "Told the Truth" Treatment Effects

	Send Difference (t-statistic)	**Return $7** Difference (t-statistic)
Full Sample (H1)		
LT - TT	-0.171 (-2.40)**	-0.222 (-3.09)***
First Round Accepters (H2)		
(LT-A) – (TT-A)	-0.181 (-2.21)**	-0.271 (-3.32)***
First Round Rejecters		
(LT-R) – (TT-R)	-0.099 (-0.70)	-0.051 (-0.34)
Anticipators (Diff-in-Diff) (H3)		
[(LT-R) – (TT-A)] – [(C-R) – (C-A)]	-0.341 (-2.13)**	-0.342 (-2.06)**
Wrong Anticipators (Diff-in-Diff)		
[(LT-A) – (TT-R)] – [(C-A) – (C-R)]	0.060 (0.34)	0.021 (0.12)

*,**,*** Significant at 10%, 5%, 1% (two-sided). t statistics (in parentheses) are constructed as described in note 28.
TT = "Told the Truth", LT = "Lied To", A = first round accepter, R = first round rejecter.

Table 5. Mood Change

	Overall	Control	Lied To	Told the Truth	Difference LT-TT (t-stat)	D-in-D LT-TT, Pos. Δ – Neg. Δ (t-stat)
Full Sample, N →	263	80	90	93		
% Positive Mood Change	0.106	0.050	0.078	0.183	-0.105 (-2.14)**	
% Negative Mood Change	0.118	0.112	0.156	0.083	0.069 (1.45)	-0.174 (-2.54)**

** Significant at 5% (two-sided). t statistics (in parentheses) are constructed as described in note 27.

Table 6. Regressions: Estimated "Lied To" (vs. "Told the Truth") Effect

	Model			
	(1) OLS	(2) OLS	(3) OLS	(4) Probit
Sample and Dependent Variable ↓	Marg. Eff. (t-stat)	Marg. Eff. (t-stat)	Marg. Eff. (t-stat)	Marg. Eff. (z-stat)
Full Sample (H1)				
Send	-0.179 (-2.51)**	-0.169 (-2.34)**	-0.186 (-2.44)**	-0.186 (-2.44)**
Return $7	-0.253 (-3.55)***	-0.236 (-3.27)***	-0.236 (-3.04)***	-0.250 (-3.06)***
	N=260	N=258	N=257	N=257
First Round Accepters (H2)				
Send	-0.191 (-2.30)**	-0188 (-2.23)**	-0.178 (-2.00)**	-0.178 (-2.00)**
Return $7	-0.310 (-3.90)***	-0.299 (-3.69)***	-0.272 (-3.13)***	-0.297 (-3.21)***
	N=193	N=192	N=191	N=191
First Round Rejecters				
Send	-0.100 (-0.66)	-0.084 (-0.54)	-0.202 (-1.19)	-0.207 (-1.30)
Return $7	-0.101 (-0.65)	-0.100 (-0.65)	-0.166 (-0.97)	-0.185 (-1.05)
	N=67	N=66	N=66	N=66
Anticipators, Diff-in-Diff, Full Sample (H3)				
	Diff-in-Diff (p-value)	Diff-in-Diff (p-value)	Diff-in-Diff (p-value)	Diff-in-Diff (p-value)
Send	-0.355 (0.030)**	-0.331 (0.046)**	-0.387 (0.021)**	-0.413 (0.021)**
Return $7	-0.344 (0.042)**	-0.311 (0.067)*	-0.327 (0.063)*	-0.336 (0.066)*
	N=260	N=258	N=257	N=257
Model Includes:				
Course Effects	Yes	Yes	Yes	Yes
Male Gender	Yes	Yes	Yes	Yes
Initial Mood	No	Yes	Yes	Yes
Mood Change	No	No	Yes	Yes

Heteroskedasticity-robust t-statistics (p-values) in parentheses. Initial mood is measured by (0,1) indicators for low mood ("bad," "down" or "so-so") and high mood ("very good" or "great"). Mood change is measured by (0,1) indicators for positive mood change and negative mood change, distinguished by "accepter"/"rejecter" in the full sample regression. p-values in the bottom ("anticipator") panel are for robust F statistics (χ^2 for probits) on linear restrictions of zero difference-in-difference for correct anticipators, [(LT-R)-(TT-A)] – [(C-R)-(C-A)]. Probit marginal effects are evaluated at means. *,**,*** Significant at 10%, 5%, 1% (two-sided).

Table 7. Normative Interpretations of Lies

	Statement: A false message is:			
	Not really lying, just being rational	Trying to trick/deceive Receiver	Not the right thing to do	Prediction
	S1	S2	S3	Q4
Percentages	**% Disagree**	**% Agree**	**% Agree**	**Predicted % Agree with S3**
Overall %	0.644	0.841	0.703	0.676
t statistic (%> 0.50)	4.89***	15.20***	7.25***	11.75***
N	264	265	266	266
Estimated LT-TT Treatment Effects for Subjects Who →	**Disagree w/ S1**	**Agree w/ S2**	**Agree w/ S3**	**Predict % Agree w/ S3 > 55%**
	Difference (t-stat)	Difference (t-stat)	Difference (t-stat)	Difference (t-stat)
Send	-0.103 (-1.08)	-0.179 (-2.06)**	-0.237 (-2.68)***	-0.252 (-2.85)***
Return $7	-0.277 (-3.01)***	-0.251 (-2.95)***	-0.246 (-2.64)***	-0.222 (-2.42)**
N	164	215	181	192

,* Significant at 5%, 1% (two-sided). LT-TT treatment effects are obtained from OLS regressions that control for course effects, gender, initial mood, and mood change. Robust t-statistics in parentheses.

Table 8. Subject Beliefs about Trust Behavior

	Participant Predictions	LT-TT Difference		
		Overall	For Accepters	For Anticipators (Diff-in-Diff)
	(1) Mean (Std. Dev.)	(2) Difference (t-stat)	(3) Difference (t-stat)	(4) Diff-in-Diff (p-value)
Q1: % Send Belief	43.27 (23.15)	-8.895 (-2.41)**	-3.660 (-0.85)	-18.636 (p=0.005)***
Q2: % Return $7 Belief	47.63 (27.54)	-8.465 (-1.89)*	-9.453 (-1.81)*	-9.247 (p=0.318)

*,**,*** Significant at 10%, 5%, 1% (two-sided). LT-TT treatment effects are obtained from OLS regressions that control for course effects, gender, initial mood, and mood change. Robust t-statistics (p-values) in parentheses.

Appendix Tables

Table A5. Mood Change for First Round Accepters and Rejecters

	Overall	Control	Lied To	Told the Truth	Difference LT-TT (t-stat)	D-in-D LT-TT, Pos. Δ – Neg. Δ (t-stat)
Accepters, N →	196	57	65	74		
% Positive Mood Change	0.117	0.053	0.046	0.230	-0.184 (-3.31)***	
% Negative Mood Change	0.122	0.123	0.215	0.040	0.175 (3.13)***	-0.359 (-4.55)***
Rejecters, N →	67	23	25	19		
% Positive Mood Change	0.075	0.043	0.160	0.000	0.160 (2.18)**	
% Negative Mood Change	0.104	0.087	0.000	0.263	-0.263 (-2.60)***	0.423 (3.39)***

,* Significant at 5%, 1% (two-sided).

Table A7. Normative Interpretations of Lies by Treatment

Percentages for ↓	Statement			Prediction
	S1	S2	S3	Q4
Control	% Disagree	% Agree	% Agree	Predicted % Agree with S3
Percent	0.587	0.864	0.679	0.658
t statistic (%> 0.50)	3.41***	9.57***	3.45***	5.43***
N	80	81	81	81
Lied To				
Percent	0.692	0.791	0.670	0.668
t statistic (%> 0.50)	3.97***	6.83***	3.46***	6.61***
N	91	91	91	91
Told the Truth				
Percent	0.645	0.871	0.755	0.698
t statistic (%> 0.50)	2.93***	10.67***	5.76***	8.37***
N	93	93	94	94
LT – TT				
Difference	0.047	-0.080	-0.085	-0.030
t-statistic	(0.68)	(-1.45)	(-1.28)	(-0.86)

*** Significant at 1% (two-sided).