Do Lies Erode Trust?

Glynis Gawn* Robert Innes†

May 2015

Abstract

Does honesty (vs. dishonesty) promote trust and trustworthiness? We investigate the effect of being lied to (or told the truth) in a Gneezy (2005) deception game on behavior in a subsequent trust game with different players. Treatment effects are decomposed between the impacts of being "burned" by a low payoff in the deception game, mood change, and the specific experience of a lie. We find that the specific experience of being lied to significantly erodes trust, trustworthiness, and the use of communication to promote trust. However, the experience effect on trustworthiness occurs only for subjects who are burned. Our results suggest that being at the receiving end of a norm violation (a lie) can relax moral preferences of the recipient, even in unrelated interactions with other people and controlling for correlated effects of the experience on payoffs, mood, and the perceived propensity for honesty in the subject pool.

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

*U.C. Merced and U.C. Berkeley. Email: ggawn@ucmerced.edu.

†County Bank Professor of Economics, U.C. Merced. Email: rinnes@ucmerced.edu. Correspondence: School of Social Sciences, Humanities and Arts, U.C. Merced, CA 95343. Fax: (209) 228-4007, Phone: (209) 228-4872.

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 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., 2008a), 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.

— 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 (Gneezy, 2005; 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 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 – by controlling for 'burns', mood, and overall propensities for honesty. Prior work demonstrates that 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

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

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. Prior work also documents that mood can affect behavior in a trust interaction (e.g., see Capra, 2004; Kirchsteiger et al., 2006). 'Burns' can alter behavior due to effects on mood, preferences, and perceptions of procedural justice (Sánchez-Pagés and Vorsatz, 2007, 2009). These general (burn/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. To identify an intrinsic effect of our treatments therefore requires controlling for their effects on payoffs — a key feature of our design.

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., 2008b; 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 some key 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 Related Literature and Conceptual Context

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; Hurkens and Kartik, 2009) 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), a norm of honesty (Pruckner and Sausgruber, 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]

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 the impact of receiving a lie on trust interactions with different players.

A number of other key papers study effects of receiving a lie, each with a different focus than ours. 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; and distinguishing between experience and 'burned' effects of being lied to — a key objective of ours — is not possible.

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

[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. (This is a small subset of the literature, and we apologize to authors of many key papers omitted here.)

Several other papers consider effects of receiving a lie 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 more lying behavior, they like and believe their 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, even after multiple rounds of play; these effects may reflect 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 'burned' effect. Sánchez-Pagés 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.[6] While these interesting results come closest to distinguishing between intrinsic experience and 'burned' effects in 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.,

[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]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.

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.

While there are potentially compelling contextual differences between all of this work and our study, an arguably more fundamental distinction is at the heart of our experiments.[7] Studies of generalized indirect reciprocity (including ours) 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. 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 arguably reflect a not-burned (good payoff) situation and signal a norm of trust and an expectation of trustworthiness.

Why do we care about the specific experience effects of lies? We are interested generally in whether being at the receiving end of a norm violation relaxes one's own moral preferences.[8] Specifically, does the experience of a lie relax the desire to reciprocate trust (with trustworthiness) and deplete trust as a result? To address this question — and understand the role of norm violations in propagating immorality and distrust — requires the identification of an intrinsic experience effect, controlling for other forces at play. You can be

[7]The treatments in prior work (for example, 'unfair'in dictator games vs. 'lied to'in our context) and/or outcomes (fairness or cheating vs. trust in our context) are very different. For example, the 'selfishness'exhibited in dictator games is sometimes heralded (by economists in particular) for promoting effort and innovation that are central to successful market economies; in other contexts, it is derided as an impediment to cooperative relationships. In contrast, dishonesty and corruption are consistently scorned by churches, community leaders, and even economists for impairing economic progress. List (2007) and Bardsley (2008) find that modest framing differences in dictator games can have significant effects on behavior, let alone more profound variations in the structure of an experiment.

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

unlucky and have a low payoff, which affects mood and behavior. You can receive a signal of an altered norm, and your behavior may change as a result. If the modeled (downstream) interaction is with the same player as the (upstream) treatment interaction, then reactions may be symptoms of reciprocal preferences and not of treatment effects on preferences.

We find that, separate from these other forces, the experience of a lie alters social preferences. 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), experience can alter the weight on reciprocation or disappointment. In a model of spite (Levine, 1998), experience may alter the value of a superior allocation. In all cases, the central content of the experience we are interested in is the norm violation (receiving a lie) or norm compliance (receiving a truth), separate from other symptoms of the treatment that might explain behavior (payoffs, mood, norms, direct reciprocity).

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.

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 C:\$6 to the Sender and \$3 to the Receiver.

Option D:\$4 to the Sender and \$6 to the Receiver.

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

Message C: Option C will earn you (the Receiver) more money than Option D.

Message D: Option D will earn you (the Receiver) more money than Option C.

Based only on the Message chosen 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 C is better for the Receiver and sometimes Option D). Receivers are never told the dollar amounts in the two options, but are told that one of the two is better for the Receiver and the other is better for the Sender.

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 summarizes our design, as described below, with a game tree.

3.2 *The Treatments*

Receivers are randomly assigned to three Treatment groups. The first group is the set of Control subjects who are exposed only to common information about the Deception game — that is, information that is given to all Receivers. The purpose of this information is to control for subject beliefs about behavior in the Deception experiment. The specific information given to all subjects, after their decisions in the Deception game (Experiment 1) are made, is as follows:[9]

> "In Experiment 1, roughly 5 out of 10 Senders TOLD THE TRUTH and 5 out of 10 Senders LIED."

This statement is based on the Sender session of our experiment, conducted prior to the Receiver sessions. The precise percentage of truthful Senders was 47.8 percent. Each subject in the second and third Treatment groups is told whether his or her own matched Sender *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.

[9]Based on results form Gneezy's (2005) experiments, where 78 percent of Receivers followed their Sender recommendations, all subjects are also told the following (after completion of Experiment 1):
"In prior sessions of Experiment 1, roughly 8 out of 10 Receivers chose the option recommended by their Senders." This added information helps to provide a complete picture of overall behavior in the Gneezy game.

How does being lied to (or being told the truth) affect (i) ones willingness to trust, (ii) ones trustworthiness, and (iii) the effect of communication in promoting trust?

In studying these questions, we note that treated subjects are distinguished not only by the Treatment information (whether they were lied to, for example) but also by their decisions in the Deception game. Subjects who *accepted* their Sender recommendations in Experiment 1 are hurt by a lie (and helped by a truthful message); conversely, subjects who *rejected* their Sender recommendations are helped by a lie, in the sense that they earn a higher payoff in Experiment 1. These distinctions will be crucial in the analysis of our experimental outcomes as we seek to disentangle effects of (i) specific experience of being lied to (or told the truth), (ii) being burned in the Deception game, and (iii) inherent differences between 'accepters' and 'rejecters'.

Random assignment to Treatments is ensured by random matching at the start of the experiment. Each questionnaire identifies a participant by the registration number, which in turn determines the Treatment group (with the correspondence known only by the experiment manager).[10] While the assignment of registration numbers to Treatments is determined a priori, assignment of registration numbers to subjects is purely random.

3.3 *The Trust Game*

After receiving the Treatments, subjects participate in a second experiment. Here, each subject 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

[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. After turning in their deception game decisions, Receivers are given an information sheet. Control subjects (with M identifiers, for example) collected their sheet at one 'station ' to which they were 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 one of two other 'stations ', 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 was made to the alphabetical identifier. On the information sheet, the LT and TT treatment subjects are told:
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.

that their matched player is a different person than their Sender from the first (Deception) experiment.

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.

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

> **Option A.** Return $7 to the Sender, so that the Returner receives $5 and the Sender receives $7.
>
> **Option B.** 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 A indicates trustworthiness. Table 1 summarizes the payments.

Table 1

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

Before the Returner decides which Option to choose, he or she can deliver a message to the Sender. The message is:[11]

[11] Our Messages take a 'bare promise' form as in Charness and Dufwenberg (CD, 2010). CD find that the 'bare promise' communication does not promote trust (vs. no communication) but promotes trustworthiness (evidence of lying aversion). As all of our subjects play with communication, our experiments do not speak to the effect of communication per se, but rather to the effect of our treatments on communication.

MESSAGE A: I am going to choose Option A.

Alternately the Returner can choose to send NO MESSAGE. Returners are told that a decision to send Message A does not preclude them from choosing Option B. The Sender decisions are elicited using the strategy method: Subjects are asked to make a choice (KEEP or SEND) for each of the two possibilities, if he or she receives Message A or No Message. Payments are then determined (by Table 1) according to the decision made by the Sender for the actual Message that was sent (Message A or No Message) and, if the Sender chooses SEND, the Returners choice of Option (Option A or Option B). Option labels are again varied between subjects (sometimes Option A is generous, as above, and sometimes stingy).

In the experiment, participants make decisions in both roles. Each matched pair is paid according to one player's decision as Sender and the other player's decision as Returner, with the allocation of roles determined by a coin flip after the experiment is completed. 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 it is the one for which you will be paid.' 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 once, whereas many experiments have participants make the same decision repeatedly. We use the strategy method for the Sender and simultaneously for the Returner (who answers contingent on a SEND decision), rather than a direct response approach.[13] We do not believe that these

[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 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.

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. Mood effects are likely to be driven by whether a subject is burned or not in the Deception experiment. We control for 'burns' in our analysis and also seek a direct measure of 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 *Logistics*

The Receiver experiment was conducted in four one-shot sessions in Economics and Sociology classes at the University of California, Merced, and Cal State East Bay.[14] All subject responses were completely anonymous, with student participants identified for payment by registration numbers. All three Treatments were conducted in all classes, resulting in a sample of 204 subjects. Sixty subjects were exposed to the Control Treatment; 72 subjects were exposed to the Lied To Treatment; and 72 subjects were exposed to the Told the Truth Treatment.

4 Results

4.1 *Baseline Results*

Tables 2 and 3 describe broad results from our experiment. Table 2 presents proportions of subjects in total, and by treatment group (Control, Lied To, and Told the Truth), who made various decisions in the trust game. The decisions include: (1) Send1, whether to Send / trust when receiving a Message from the Returner indicating that s/he (the Re-

[14]The three Economics sessions were upper division and the one Sociology session was lower division. The Sender side of the experiment was conducted in a large introductory Economics class at U.C. Merced that minimized overlap with the Receiver sessions. Course rosters were used to ensure that no subject participated twice. Subjects were paid one week after the experiment was completed; to obtain payment, each student produced a tag with the registration number that was attached to the original questionnaire.

turner) intends to choose the generous option (Option A in Table 1); (2) Send2, whether to Send/trust when the Returner sends no Message; (3) OptGen, whether to choose the generous option (the Returner's decision), which we refer to as the trustworthy decision (following standard nomenclature); (4) MessGen, whether to Send a Message indicating selection of the generous option (again the Returner's decision); and (5) various combined choices of the Returner, including a Deceitful strategy (sending the Message, but choosing the ungenerous option, Deceit), an untrustworthy but not deceitful combination (not sending the Message and choosing the ungenerous option, UTBND), a trustworthy and truthful strategy (sending the Message and choosing the generous option, TWTruth), and a trustworthy strategy without communication (choosing the generous option and no Messaage, TWBNM).

Table 3 presents z-statistics for the respective differences between choices of the three treatment groups: (1) Lied To versus Control (column (1)), (2) Told the Truth versus Control (column (2)), and (3) Lied To versus Told the Truth (column (3)). The Table reveals that Lied To subjects were significantly less trusting (in terms of Send1); less inclined to send a Message; less likely to be trustworthy and truthful; and more likely to be trustworthy without sending a Message, all relative to both Control subjects and subjects who were Told the Truth. Lied To subjects were also significantly less likely to be trustworthy overall (by choosing the generous option), and more likely to be untrustworthy but not deceitful, relative to subjects who were Told the Truth. The absolute magnitudes of these differences are noteworthy. For example, 43 percent of Lied To (LT) subjects were trusting (Send1) compared with 61 percent of subjects who were Told the Truth (TT); 46 percent of LT subjects chose the generous option, compared with 62.5 percent of TT subjects; and 53 percent of LT subjects sent a Message, compared with over 76 percent for the TT group.

The bottom panels of Tables 2 and 3 present corresponding statistics for (i) the fraction of subjects who accepted/followed their Sender recommendations in the deception game, (ii) the initial (pre-experiment) mood report of participants (on a scale of zero to five, from bad to great), and (iii) the fraction who (after the deception game was complete and the treatments received) reported mood increases and mood decreases, respectively. A check for random assignment of our treatments is provided by comparison of the accept/follow decisions and initial moods of the subjects across treatment groups. If we have random

assignment, there should be no significant differences between these indicators across the treatments, and indeed, Table 3 reveals no significant differences. Overall, approximately 62 percent of our student participants chose to accept/follow their Sender recommendation in the Deception game, with only slight variation from one treatment group to another.

Treatments did, however, affect mood changes in predictable directions. LT subjects were significantly more likely to experience mood decreases relative to either the Control or TT participants. These results both provide some validation of our mood measure and mean that some of the treatment effects on the trust game (LT versus Control and TT) could be attributable to resulting impacts on mood. We return to this issue in a moment.

Table 4 supplements Table 3 by reporting Probit regression results for Sender and Returner decisions, controlling for the subjects gender, initial mood, and fixed course effects. While the added correlates increase precision in estimated treatment effects, the broad conclusions of Tables 2-3 are upheld. For example, the Lied To treatment is estimated to reduce the probability of trust (Send1) by 19.3 percent, and the probability of a trustworthy and truthful strategy by 19.8 percent, both effects statistically significant.

Arguably surprising (and related) features of our baseline results are the large fraction of subjects who choose to Send when no Message is received (Send2) and the non-negligible fraction of subjects who are trustworthy but send no Message indicating their choice (TWBNM). In all treatments, over a third of subjects elect to Send when no Message is received.[15] On the Returner side, 11.8 percent of subjects choose TWBNM, which is roughly one-third of subjects who send no Message; corresponding fractions are highest for Lied To participants (19.4 percent of whom choose TWBNM, out of 47.2 percent who send no Message), next highest for participants who are Told the Truth (9.7 percent choosing TWBNM out of 23.6 percent who send no Message), and lowest for Control participants (5 percent choosing TWBNM out of 31.7 percent who send no Message). Charness and Dufwenberg (2010) observe similar fractions in a similar experiment, but with different payoffs and much smaller subject numbers than we have; in their experiment, two of seven Senders who received no

[15]On the sender side in our experiment, a risk neutral subject purely interested in his own payoffs would choose the Send2 strategy only if the probability of a generous Returner (given that no Message is sent) is 40 percent or higher. This condition is violated in the Control group and in our overall sample (the relevant benchmark given random matching across all subject participants).

Message chose to Send (Send2), and three of seven Returners who sent no Message chose the generous option (TWBNM). We find that this phenomenon is more general and cannot be explained by randomness in participant choices. Perhaps the TWBNM strategy is motivated by psychic rewards to perceived acts of virtue untainted by a self-interested Message. If so, a reduced saliency of Messages might be expected to tip this calculus in favor of the TWBNM strategy, as we observe for the Lied To subjects.

4.2 *Decomposing Treatment Effects*

From our experiment, we are interested not only in identifying broad effects of our treatments — exposure to dishonesty or honesty — on behavior in trust relationships, but also the impact of this specific experience, as separate and distinct from mood effects and impacts of being burned or not in the first experiment (which can also affect mood and behavior). Our baseline comparisons (in Tables 2-4 above) potentially conflate these phenomena. When a subject accepts his Sender recommendation in our Deception game (Experiment 1), he is hurt / burned when Lied To and benefited / not-burned when Told the Truth; conversely, when a subject rejects his Sender recommendation, he is not burned when Lied To and burned when Told the Truth. Now, if acceptance and rejection of Sender recommendations were equi-probable, there would be no differences in propensities to be burned or not burned across the two (LT and TT) treatment groups; in this case, cross-group differences could not be explained by a burned composition effect. However, in our experiment, roughly 62 percent of participants accepted their Sender recommendations, meaning a much higher fraction of burned subjects in the LT treatments than in the TT treatments. Our baseline results could therefore be explained by treatment effects on being burned, rather than a pure experience effect of being Lied To. For example, Lied To subjects may be less trusting because they are more likely to have been burned.

To disentangle these effects, we first break down our subject decisions by both treatment group and Acceptance / Rejection decisions from Experiment 1. The decomposed summary statistics from the experiment are given in Table 5. In principle, one way to net out burn effects would be to compare LT Accepters to TT Rejecters, both of whom are burned, and LT Rejecters to TT Accepters, neither of whom are burned. However, this comparison conflates potentially different drivers of Receiver Acceptance / Rejection decisions; accepters may be

different types of people than rejecters. Indeed, the last column of Table 5 reports difference statistics for behavior of Control Accepters and Control Rejecters in our experiment. Because all Controls are equally likely to have been lied to or told the truth, there is no differential burned effect for Control Accepters vs. Rejecters. However, there are several key differences in behavior. The Control Accepters are significantly less responsive to communication in their trust decision, significantly more likely to choose the generous option, significantly less likely to be deceitful, and significantly more likely to be trustworthy without sending a Message. In sum, these statistics indicate that Accepters come from a different population than Rejecters, meaning that the comparisons proposed above would conflate the experience effect of being Lied To (vs. Told the Truth) with differences between Accepters and Rejecters.

We overcome this confound by constructing difference-in-difference statistics that exploit the Control subjects to adjust LT-versus-TT differences for burned and not-burned subjects, respectively; this is done by netting out corresponding differences between Control Accepters and Rejecters. For burned subjects, the difference-in-difference takes the difference between LT Accepters (LTA) and TT Rejecters (TTR), and subtracts out the corresponding difference between Control Accepters (CA) and Control Rejecters (CR). This difference-in-difference gives us a pure experience (vs. burned) effect of being Lied To (vs. Told the Truth). Similarly, for not-burned subjects, the difference-in-difference compares LT Rejecters (LTR) to TT Accepters (TTA), and subtracts out the corresponding difference between Control Rejecters and Control Accepters. Parallel difference-in-difference statistics give the pure burned effect for LT subjects, (LTA-LTR)-(CA-CR), and for TT subjects, (TTR-TTA)-(CR-CA).

Table 6 presents a first set of these decompositions. The first two columns give the pure Lied To (vs. TT) experience effect for burned and not-burned subjects, respectively; the third and fourth columns give pure burned effects for TT and LT subjects, respectively. z-statistics for the difference-in-differences are given in parentheses.[16] Columns (1) and (3) (and columns (2) and (4)) add up to the joint LT (vs. TT) effect for accepters (LTA-TTA), combining the experience and burned effects of the different treatments; this joint effect is

[16]The z-statistics are calculated as $z = D/se$, where D=difference in difference=$(p_1 - p_2) - (p_3 - p_4)$ and $se = [\sum_{i=4}^{4} v_i/n_i]^{1/2}$ where $v_i = p_i(1 - p_i)$, p_i=proportion in sample i, and n_i=size of sample i. For Send1-Send2, sample variances are used for the variance estimates v_i.

presented in column (5).

At the bottom of Table 6, we find that the propensity for a negative mood change is significantly raised by the LT (vs. TT) experience (for the not burned) and by the burned experience (for the TT subjects). The propensity for a positive mood change is significantly reduced by the burned experience (for the LT subjects). Being burned thus worsens our subjects' moods. Being Lied To also worsens mood, at least for those not burned. Because of the latter effect, we want to construct difference-in-difference statistics for trust outcomes that control for mood changes directly.

Table 7 presents generalized difference-in-difference statistics for pure LT (vs. TT) experience and pure burned effects, respectively, that control for gender, course effects, initial mood, and mood changes (positive and negative). These statistics are constructed from robust tests of coefficient differences in linear probability (OLS) estimations; p-values for the test statistics are reported in parentheses.

4.3 *Main Results*

Tables 6 and 7 reveal broadly similar experience and burned effects, and give us the main conclusions from our experiment: First, we find that *being Lied To (versus Told the Truth) erodes trust* both for burned and not-burned subjects. However, for the not burned, trust is eroded when a Message is sent (Send1, column (2)), whereas for the burned, trust is eroded when a Message is not sent (Send2, column (1)). The LT experience effects are large. For the not-burned, the LT experience reduces the propensity for trust Send1 by an estimated 33.5 percent (Table 7), compared with an average rate of trust for Control subjects of 58.3 percent (Table 2). For the burned, the LT experience reduces the propensity for trust Send2 by an estimated 39.7 percent (Table 7), compared with a Control subject propensity of 36.8 percent (Table 2). These numbers capture the intrinsic effects of lies that we discussed at the start of the paper, separate from any treatment effects on mood and/or being burned or not in Experiment 1.

We expect effects of being burned to be different for TT and LT subjects. TT subjects are burned when not following their Sender recommendation in Experiment 1; we expect being burned to motivate different (more trusting) behavior in the trust game and/or to make the signal of truthfulness more salient, again favoring more trusting choices. Conversely,

LT subjects are burned when following their Sender recommendations in Experiment 1; we therefore expect being burned to motivate less trusting behavior in the trust game. For Send1, however, we find no significant burned effect, leading to a joint (experience and burned) effect of the LT treatment that is negative and significant (column (5), Tables 6-7). For Send2, we find a significant positive burned effect on TT subjects, consistent with expectations, but no burned effect on LT subjects. The joint effect combines the negative experience effect (of LT) on Send2 (column (1), Tables 6-7) with the positive burned effect (for the TT subjects, column (3), Tables 6-7), for a net null effect.

Second, *being Lied To (versus Told the Truth) and being burned interact to erode trustworthiness*. The experience effect of being Lied To (for the burned), and the burned effect (for the Lied To), are to reduce trustworthiness by a statistically and economically significant fraction. The LT experience reduces the propensity for overall trustworthiness (OptGen) by an estimated 51.5 percent and the propensity for both truth and trustworthiness (TWTruth) by an estimated 45.1 percent (Table 7, column (1)). The burned effect (for the Lied To) reduces overall trustworthiness (OptGen) by an estimated 57.2 percent and TWTruth by an estimated 35.8 percent (Table 7, column (4)). However, we find no significant LT experience effect on trustworthiness for the not-burned, and no significant burned effect for the TT subjects. Hence, being Lied To and being burned each reduce trustworthiness in our experiment, but only when the other is present.

This general conclusion is reinforced by two more nuanced results. Being Lied to (for the burned) is estimated to raise the likelihood of untrustworthiness with no deceit (UTBND); and being burned (for the Lied To) is estimated to lower the likelihood of trustworthy behavior with no Message (TWBNM). Interestingly, however, we find no significant Lied To effects on the propensity for Deceit (an untrustworthy choice together with a deceitful Message indicating the opposite).

Third, the propensity to send a Message is reduced both by the pure experience effect of being Lied To (versus Told the Truth) and by being burned in Experiment 1 (columns (1)-(4), Tables 6-7). However, none of these effects is statistically significant individually and the estimated impact of being burned is particularly small (see p-values in columns (3)-(4) of Table 7). Combining Lied To and burned effects on the Accepters (column (5) of Tables

6-7) therefore reflects primarily the experience effect. And the combined effect — which is measured with more precision — is statistically significant and negative. We conclude that *the Lied To experience reduces communication*. For example, the Lied To experience, for burned subjects, reduces the fraction of Returners who send a Message by 26.8 percent (Table 7, column (1)), compared with an overall propensity to send Messages of 68.3 percent among Control participants (Table 2). The reduced reliance on communication to promote trust generally contributes to reduced trustworthiness, as described above.

In summary, we find that the pure experience effect of being Lied To (versus Told the Truth) erodes trust, trustworthiness, and communication in our experiment. These effects are distinct from (and control for) treatment impacts on mood and being burned.

4.4 *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. Given communication in our experiment, this question is well posed only when conditioned on the receipt of a Message. In order to construct this modified STZ (2013) measure of trust, we used an incentive compatible approach to elicit this belief, along with three others, from subjects in our experiment.

Specifically, we asked subjects to predict four outcomes from the experiment, paying $1 for each prediction that was within 5 percent (plus or minus) of the true percentage (using 5 percentage point bands). The outcomes for which we solicited predictions are:

Q1. The fraction of Senders who Send when receiving a Message (Send1).

Q2. The fraction of Returners who, if Sending a Message, choose the generous option.

Q3. The fraction of Returners who send a Message.

Q4. The fraction of participants who indicate (in a separate yes or no question) that a Returner should choose the generous option if sending a Message.

Q2 gives the STZ measure of trust for our game. Table 8 provides summary and difference-in-difference statistics for the four beliefs, as well as whether subjects think a Returner should

choose the generous option if sending a Message (Q5, 1 for yes, 0 for not necessarily).

Broadly, we find two main differences between subject answers on the belief and norm questions across the treatment groups (LT, TT, and Control). Being Lied To has a significant negative effect on the perceived likelihood of trust Send1 (Q1) and the perceived norm on whether Returners should be trustworthy when promising to be so (Q4). These results might reflect false consensus effects — that is, subject beliefs that conform to the subjects' own choices (Ross, et al., 1977; Ellingsen, et al., 2010; Butler et al., 2015).[17]

Decomposing the treatment effects further, with difference-in-difference statistics, we find only one significant effect. The Lied To experience, for burned subjects, significantly reduces the predicted frequency with which Returners will be generous when sending the Message (Q2). In other words, we find a negative Lied To effect on the STZ measure of trust. Note, however, that we find no significant Lied To experience effects on beliefs and norms for the not-burned. Hence, the experience effects that we identify in column (2) of Tables 6-7 cannot be attributed to treatment effects on beliefs, at least not the ones that we measure. Most importantly, the significant Lied To experience effect in eroding trust (Send1, for the not-burned) does not appear to be attributable to beliefs.

5 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.

A key feature of the analysis is the identification of a specific experience effect of the Lied To and Told the Truth treatments, controlling for the impact of being burned or not, related effects on mood, and overall Sender propensities for honesty. Separate from everything else,

[17] We performed all of the difference-in-difference calculations of Table 7, adding in the belief question Q4 to the corresponding OLS regressions. Because the beliefs are potentially endogenous (and we have no instruments with which to identify them), we do not report these estimations here. However, the results are available upon request and consistent with those reported in Table 7 above.

the 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 choices.

6 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.

Bardsley, N. (2008). "Dictator game giving: Altruism or artifact?" *Experimental Economics*, 11, 122-133.

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.

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*, in press.

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.

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.

List, J. (2007). "On the interpretation of giving in dictator games." *Journal of Political Economy*, 115, 482-493.

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.

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.

Pruckner, G. and R. Sausgruber. 2013. "Honesty on the Streets: A Field Study on Newspaper Purchasing." *Journal of the European Economic Association*, 11(3), 661-79.

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.

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.

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.

Figure 1A: Gneezy Game and Receiver Treatment

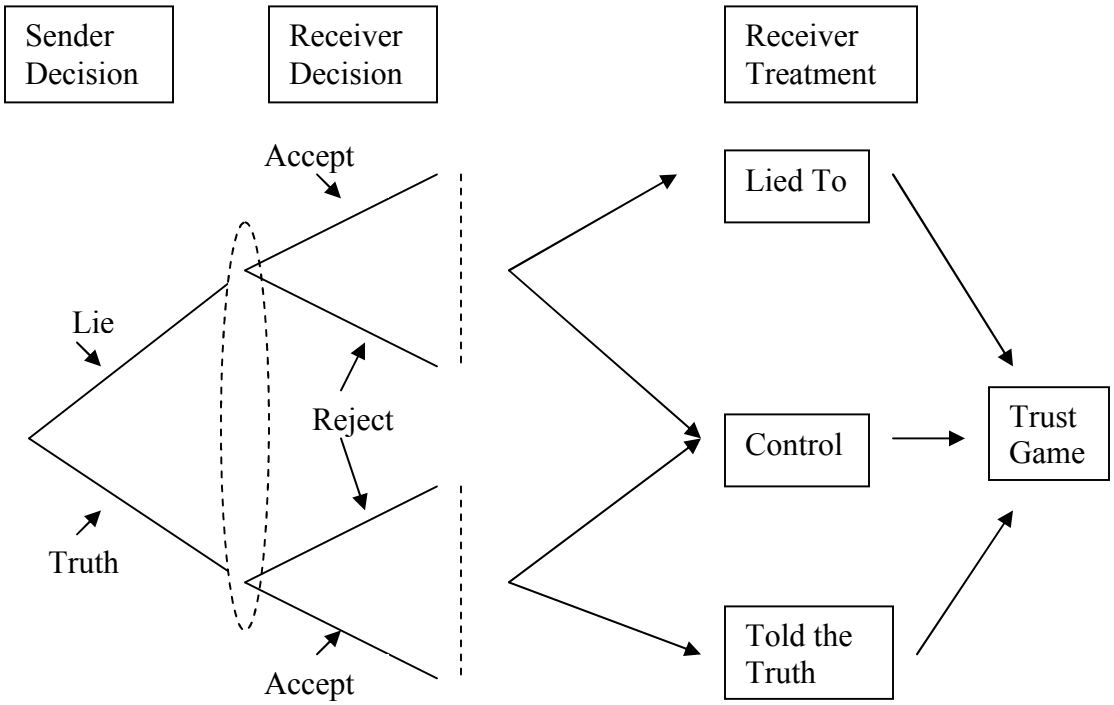

Figure 1B: Trust Game (Gneezy Receivers Play Both Roles)

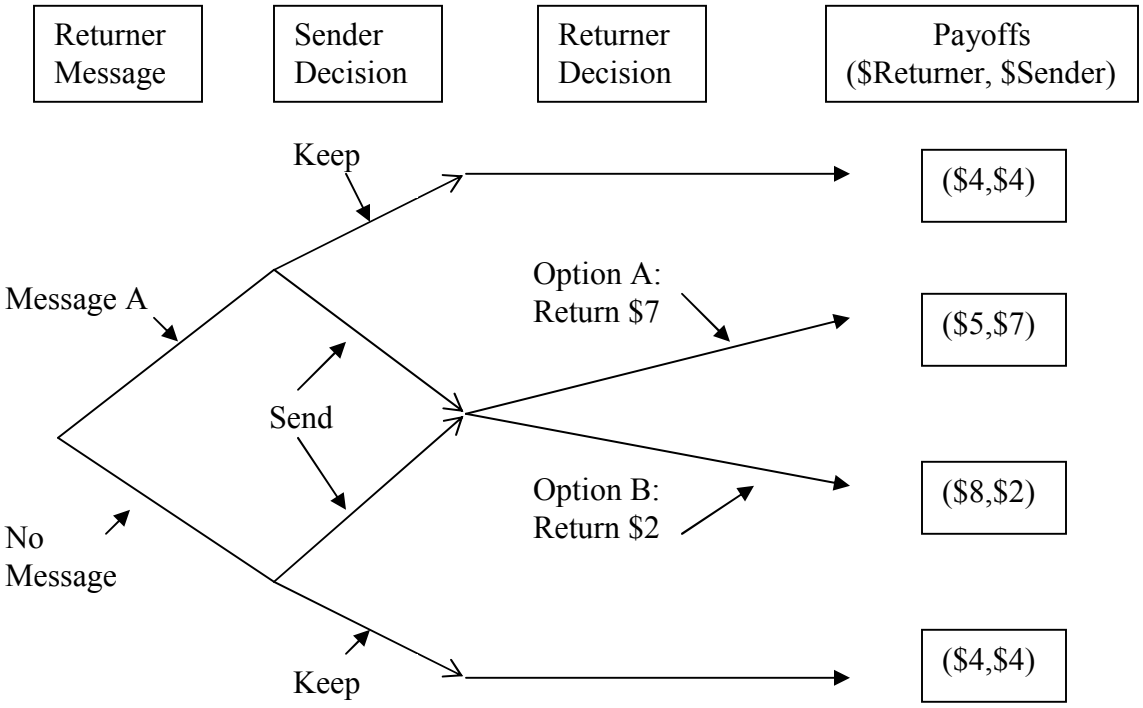

Table 2. Sample Summary Statistics

	All Observations (n=204)	Treatments		
		Control (n=60)	Lied To (n=72)	Told the Truth (n=72)
Trust				
Send1 (Trust When Message Rec'd)	0.539	0.583	0.431	0.611
Send2 (Trust When No Message Rec'd)	0.368	0.333	0.361	0.403
Send1 – Send2 (Effect of Message on Trust)	0.172	0.250	0.069	0.208
Trustworthiness				
OptGen (Trustworthy: Generous Option Chosen)	0.525	0.483	0.458	0.625
MessGen (Message Sent)	0.659	0.683	0.528	0.764
Deceitful (Message Sent, but Untrustworthy)	0.250	0.250	0.264	0.236
UTBND (Untrustworthy, but no deceit/no Message)	0.225	0.267	0.278	0.139
TWTruth (Trustworthy & Truthful Message)	0.407	0.433	0.264	0.528
TWBNM (Trustworthy, No Message)	0.118	0.050	0.194	0.097
Deception Game Decision and Mood				
Accept (in Deception Game)	0.623	0.617	0.597	0.653
Initial Mood	2.936	2.783	2.958	3.042
Positive Mood Change^{+}	0.127	0.100	0.153	0.125
Negative Mood Change^{+}	0.113	0.067	0.222	0.042

+ Positive (Negative) Mood Change = 1 if Mood Change (post-treatment minus pre-treatment) > (<) 0, 0 otherwise.

Table 3. Difference Statistics Across Treatments

	Lied To vs. Control (z-statistic)	Told the Truth vs. Control (z-statistic)	Lied To vs. Told the Truth (z-statistic)
Trust			
Send1	-1.77*	0.32	-2.20**
Send2	0.35	0.83	-0.52
Send1 – Send2	-3.05***	-0.57	-2.46**
Trustworthiness			
OptGen	-0.30	1.65	-2.04**
MessGen	-1.93*	1.03	-3.06***
Deceitful	0.19	-0.19	0.39
UTBND	0.15	-1.82*	2.08**
TWTruth	-2.17**	1.08	-3.36***
TWBNM	2.71***	1.05	1.67*
Deception Decisions and Mood			
Accept	-0.24	0.43	-0.69
Initial Mood	0.92	1.35	-0.50
Positive Mood Change	0.95	0.45	0.48
Negative Mood Change	2.70***	-0.63	3.32***

*,**,*** Significant at 10%, 5%, 1% (two-sided).

Table 4. Probit Regressions

Dependent Variable		Marginal Effect (Robust t-statistic)					LT vs. TT
		Lied To (LT)	Told the Truth (TT)	Male Gender	Initial Mood	Course Effects	Difference (p-value)[+]
Send1 (Trust When Message Rec'd)	Model 1	-0.193 (-2.117)**	-0.007 (-0.081)	0.065 (0.895)	0.089 (2.535)**	Yes	-0.186 (0.028)**
	Model 2	-0.160 (-1.823)*	0.019 (0.218)	0.072 (1.015)	No	No	-0.141 (0.032)**
Send 2 (Trust When No Message Rec'd)	Model 1	0.042 (0.481)	0.084 (0.975)	0.058 (0.842)	-0.014 (-0.431)	Yes	-0.042 (0.604)
	Model 2	0.043 (0.508)	0.083 (0.971)	0.052 (0.764)	No	No	-0.040 (0.627)
Send1-Send2 (Effect of Message on Trust)[++]	Model 1	-0.226 (-2.039)**	-0.090 (-0.838)	0.005 (0.053)	0.021 (2.328)*	Yes	-0.136 (0.205)
	Model 2	-0.201 (-1.812)*	-0.063 (-0.575)	0.018 (0.211)	No	No	-0.138 (0.197)
MessGen (Message Sent)	Model 1	-0.155 (-1.830)*	0.081 (0.961)	0.136 (1.977)**	0.010 (0.320)	Yes	-0.236 (0.004)***
	Model 2	-0.162 (-1.940)*	0.075 (0.889)	0.145 (2.152)**	No	No	-0.087 (0.003)***
OptGen (Trustworthy: Generous Option Chosen)	Model 1	-0.043 (-0.476)	0.133 (1.508)	0.059 (0.816)	0.037 (1.05)	Yes	-0.176 (0.037)**
	Model 2	-0.014 (-0.163)	0.151 (1.730)*	0.049 (0.695)	No	No	-0.165 (0.047)**
Deceitful (Message Sent, Untrustworthy)	Model 1	0.047 (0.597)	0.005 (0.066)	0.045 (0.727)	-0.039 (-1.360)	Yes	0.042 (0.573)
	Model 2	0.013 (0.168)	-0.017 (-0.221)	0.053 (0.870)	No	No	0.030 (0.684)
UTBND (Untrustworthy, No Message)	Model 1	0.004 (0.056)	-0.132 (-1.892)*	-0.101 (-1.679)*	-6.5e-05 (-0.002)	Yes	0.136 (0.047)**
	Model 2	0.004 (0.052)	-0.132 (-1.87)*	-0.101 (-1.711)*	No	No	0.136 (0.046)**
TWBNM (Trustworthy, No Message)	Model 1	0.184 (2.664)***	0.091 (1.356)	-0.026 (-0.665)	-0.008 (-0.408)	Yes	0.093 (0.098)*
	Model 2	0.200 (2.783)***	0.101 (1.457)	-0.040 (-0.998)	No	No	0.099 (0.099)*
TWTruth (Trustworthy & Truthful Message)	Model 1	-0.198 (-2.255)**	0.070 (0.807)	0.088 (1.213)	0.044 (1.245)	Yes	-0.268 (0.001)***
	Model 2	-0.179 (-2.078)**	0.087 (1.012)	0.089 (1.266)	No	No	-0.266 (0.001)***

*,**,*** Significant at 10%, 5%, 1% (two-sided).

[+] p-value for test of equal coefficients on Lied To and Told the Truth (heteroskedasticity-robust). [++]OLS for the difference between the 0-1 choices to Trust with a Message and to Trust without a Message. All other models are Probit estimations.

Table 5. Decomposed Summary Statistics Across Treatments

	Control Accept (n=37)	Control Reject (n=23)	LiedTo Accept (n=43)	LiedTo Reject (n=29)	Told the Truth Accept (n=47)	Told the Truth Reject (n=25)	Control Accept – Control Reject Difference (z-statistic)
Trust							
Send1	0.514	0.696	0.442	0.414	0.617	0.600	-0.182 (-1.44)
Send2	0.405	0.217	0.349	0.379	0.340	0.520	0.188 (1.59)
Send1 – Send2	0.108	0.478	0.093	0.034	0.277	0.080	-0.370 (-3.19)***
Trustworthiness							
OptGen	0.568	0.348	0.349	0.621	0.638	0.600	0.220 (1.71)*
MessGen	0.649	0.739	0.488	0.586	0.745	0.800	-0.090 (-0.75)
Deceitful	0.162	0.391	0.279	0.241	0.191	0.320	-0.227 (-1.93)*
UTBND	0.270	0.261	0.372	0.138	0.170	0.080	0.009 (0.08)
TWTruth	0.486	0.348	0.209	0.345	0.553	0.480	0.138 (1.08)
TWBNM	0.081	0.000	0.140	0.276	0.09	0.120	0.081 (1.81)*
Mood							
Initial Mood	2.838	2.696	2.91	3.034	3.085	2.96	0.142 (0.508)
Positive Mood Change	0.054	0.043	0.070	0.276	0.170	0.04	0.011 (0.19)
Negative Mood Change	0.135	0.044	0.302	0.103	0.000	0.120	0.091 (1.29)

*,**,*** Significant at 10%, 5%, 1% (two-sided).

Table 6. Difference-in-Difference Statistics

	Lied To Effect		**"Burned" Effect**		**Total Lied To and Burned Effect for Accepters**
	For the "Burned"	For the "Not Burned"	For the Told-Truth	For the Lied To	
	Diff-in-Diff (z-statistic)	Diff-in-Diff (z-statistic)	Diff-in-Diff (z-statistic)	Diff-in-Diff (z-statistic)	Difference (z-statistic)
	(1) (LTA-TTR) - (CA-CR)	(2) (LTR-TTA) - (CR-CA)	(3) (TTR-TTA) - (CR-CA)	(4) (LTA-LTR) - (CA-CR)	(5) LTA-TTA
Trust					
Send1	0.024 (0.13)	-0.385 (-2.25)**	-0.199 (-1.12)	0.210 (1.24)	-0.175 (-1.69)*
Send2	-0.359 (-2.68)***	0.227 (1.39)	0.368 (2.14)**	-0.218 (-1.35)	0.009 (0.08)
Send1-Send2	0.383 (1.72)*	-0.612 (-2.81)***	-0.567 (-2.61)***	0.429 (1.93)*	-0.184 (-1.16)
Trustworthiness					
OptGen	-0.471 (-2.62)***	0.202 (1.17)	0.181 (1.02)	-0.492 (-2.91)***	-0.290 (-2.87)***
MessGen	-0.221 (-1.33)	-0.249 (-1.52)	-0.035 (-0.22)	-0.007 (-0.04)	-0.256 (-2.58)***
Deceitful	0.188 (1.11)	-0.179 (-1.16)	-0.101 (-0.61)	0.267 (1.74)*	0.087 (0.98)
UTBND	0.283 (1.89)*	-0.023 (-0.16)	-0.081 (-0.57)	0.225 (1.51)	0.202 (2.20)**
TWTruth	-0.409 (-2.31)**	-0.070 (0.40)	0.065 (0.36)	-0.274 (-1.67)*	-0.344 (-3.60)***
TWBNM	-0.062 (-0.63)	0.272 (2.65)***	0.116 (1.28)	-0.217 (-2.00)**	0.054 (0.82)
Mood					
Initial Mood	-0.195 (-0.51)	0.092 (-0.23)	0.017 (0.05)	-0.270 (-0.74)	-0.178 (-0.85)
Positive Mood Change	0.019 (0.240)	0.116 (1.02)	-0.120 (-1.35)	-0.217 (-2.03)**	-0.101 (-1.49)
Negative Mood Change	0.091 (0.75)	0.195 (2.15)**	0.211 (2.16)**	0.108 0.95	0.302 (4.32)***

*,**,*** Significant at 10%, 5%, 1% (two-sided).
Legend: CA = Control Accept, CR = Control Reject, LTA = Lied To Accept, LTR = Lied To Reject, TTA = Told the Truth Accept, TTR = Told the Truth Reject.

Table 7. Difference-in-Difference Statistics Controlling for Course Effects, Gender, Initial Mood, and Mood Change

	Lied To Effect		**"Burned" Effect**		**Total Lied To and Burned Effect for Accepters**
	For the "Burned"	For the "Not Burned"	For the Told-Truth	For the Lied To	
	Diff-in-Diff (p-value)	Diff-in-Diff (p-value)	Diff-in-Diff (p-value)	Diff-in-Diff (p-value)	Difference (p-value)
	(1) (LTA-TTR) - (CA-CR)	(2) (LTR-TTA) - (CR-CA)	(3) (TTR-TTA) - (CR-CA)	(4) (LTA-LTR) - (CA-CR)	(5) LTA-TTA
Trust					
Send1	-0.031 (0.867)	-0.335 (0.062)*	-0.163 (0.379)	0.141 (0.449)	-0.194 (0.083)*
Send2	-0.397 (0.028)**	0.25 (0.139)	0.381 (0.030)**	-0.266 (0.132)	-0.016 (0.881)
Send1-Send2	0.366 (0.109)	-0.585 (0.007)***	-0.544 (0.014)**	0.407 (0.078)*	-0.178 (0.214)
Trustworthiness					
OptGen	-0.515 (0.006)***	0.279 (0.118)	0.221 (0.227)	-0.572 (.002)***	-0.294 (0.007)***
MessGen	-0.268 (0.126)	-0.214 (0.201)	-0.048 (0.772)	-0.102 (0.564)	-0.316 (0.003)***
Deceitful	0.183 (0.269)	-0.231 (0.124)	-0.159 (0.325)	0.256 (0.111)	-0.024 (0.798)
UTBND	0.332 (0.041)**	-0.048 (0.748)	-0.063 (0.677)	0.317 (0.059)*	0.269 (0.005)***
TWTruth	-0.451 (0.013)**	0.017 (0.925)	0.111 (0.228)	-0.358 (0.047)**	0.340 (0.002)***
TWBNM	-0.063 (0.523)	0.262 (0.014)**	0.111 (0.548)	-0.214 (0.052)*	0.048 (0.507)

Legend: CA = Control Accept, CR = Control Reject, LTA = Lied To Accept, LTR = Lied To Reject, TTA = Told the Truth Accept, TTR = Told the Truth Reject. Difference-in-difference (and difference) statistics are obtained from OLS regressions that include treatment dummies, course effects, gender, initial mood, positive mood change, and negative mood change. p-values are presented for heteroskedasticity-robust F statistics for linear restrictions of zero difference-in-difference (zero difference in column (5)).

Table 8. Subject Beliefs about Behavior and Norms

Participant Predictions →	Q1 (% choosing Send1)	Q2 (% choosing OptGen when choosing MessGen)	Q3 (% choosing MessGen)	Q4 (% saying "yes" on Q5)	Q5 (Norm) (1 = "yes" = *should* choose OptGen if send MessGen)
Summary Statistics: Mean (Standard Deviation)					
All Obs. (n=204)	51.005 (25.318)	47.857 (24.736)	55.025 (25.700)	53.719 (22.201)	42.365 (49.536)
Control (n=60)	55.583 (23.526)	46.583 (23.926)	56.667 (25.926)	54.915 (20.605)	55.932 (50.073)
Lied To (LT) (n=72)	45.278 (25.604)	44.577 (26.949)	53.611 (26.474)	48.333 (23.271)	33.333 (47.471)
Told Truth (TT) (n=72)	52.917 (25.739)	52.153 (22.764)	55.069 (24.972)	58.125 (21.533)	40.278 (49.390)
Difference Statistics (z-statistics)					
LT – Control	-10.305 (-2.41)**	-2.006 (-0.45)	-3.056 (-0.67)	-6.582 (-1.72)*	-22.599 (-2.66)***
TT – Control	-2.666 (-0.62)	5.570 (1.36)	-1.598 (-0.36)	3.210 (0.87)	-15.654 (-1.81)*
LT – TT	-7.639 (-1.79)*	-7.576 (-1.82)*	-1.458 (-0.34)	-9.792 (-2.62)***	-6.945 (-0.87)
Difference-in-Difference [+] (p-values)					
Lied To Effect for "Burned"	0.069 (0.994)	-17.675 (0.045)**	6.483 (0.491)	-8.318 (0.314)	-1.787 (0.924)
Lied To Effect for "Not Burned"	-9.036 (0.272)	-0.455 (0.959)	-5.605 (0.525)	-10.601 (0.162)	-11.985 (0.340)
"Burned" Effect for TT	-10.475 (0.220)	11.772 (0.158)	-12.149 (0.181)	0.792 (0.919)	-1.842 (0.920)
"Burned" Effect for LT	-1.37 (0.875)	-5.448 (0.566)	-0.061 (0.995)	3.076 (0.703)	8.356 (0.661)

*,**,*** Significant at 10%, 5%, 1% (two-sided).

[+] Difference in difference statistics are calculated from heteroskedasticity-robust OLS regressions that control for course effects, gender, initial mood, and mood changes. Lied To Effect for "Burned" = (LTA-TTR)-(CA-CR), Lied To Effect for "Not Burned" = (LTR-TTA)-(CR-CA), "Burned" Effect for TT = (TTR-TTA)-(CR-CA), "Burned" Effect for LT = (LTA-LTR)-(CA-CR).