

Language and Lies

Glynis Gawn ^{*} Robert Innes [†]

April 2018

Abstract

Does an individual's aversion to a lie depend upon the language used to communicate the lie? We adapt the Lopez-Perez & Spiegelman (2013) dot experiment to measure how a "weak" vs. "strong" message affects individuals' propensities for truthfulness when there is a monetary incentive to lie and no other person is affected by the communication. Weak messages state a fact, whereas strong statements "solemnly swear" to the fact. In our first (between-subject) experiment, strong (vs. weak) statements increase the percentage of subjects choosing to tell the truth by approximately 30 percentage points in each of three different payoff scenarios that favor lying to a different extent. Because lies increase payoffs in the experiment, the weaker aversion to weaker lies is socially advantageous. In a second (within-subject) experiment participants choose between messages of different strength and we find (1) a preference for lying with weak (vs. strong) language, and (2) a significant fraction of subjects who are willing to pay a positive amount to avoid a strong vs. weak lie. From both experiments, we conclude that our subjects tend to be intrinsically less averse to dishonesty when a lie is conveyed with weak vs. strong language.

Keywords: Deception, Language, Communication, Lying Aversion

JEL Classifications: D03

^{*}Economics, School of Social Sciences, Humanities & Arts, University of California, Merced. Email: ggawn@ucmerced.edu.

[†]Corresponding Author: Economics, School of Social Sciences, Humanities & Arts, University of California, Merced, 5200 N. Lake Rd., Merced, CA 95343. Email: rinnes@ucmerced.edu. Phone: (209) 228-7742.

1 Introduction

In the standard economic model of self-interested behavior, individuals will only be honest if their economic interests are not harmed by doing so. In this model, communication is “cheap talk” and incentives must be designed to make honesty pay and thereby elicit truthful statements and reports. However, contrary to this model, research over the past thirty years documents that communication is important in promoting coordination and trust. Moreover, recent literature shows a broad tendency for individuals to be intrinsically averse to lies, and to have a corresponding preference for truthfulness.¹ Such preferences can help to make communication meaningful in economic interchange.

If communication matters, then the specific language used in the communication may be important. How does language variation intermediate the effect of communication? Why does language variation occur and why might it be advantageous? Literatures in cognitive science and linguistics document how subtle differences in language can have surprisingly profound impacts on the perceptions of an audience (e.g., Matlock, 2012; Talmy, 2000). Key studies in economics show that “strong” promises have a greater effect in promoting trust (Charness & Dufwenberg (CD), 2006; Houser & Xiao, 2011), deterring lies (Lundquist, Ellingsen, Gribbe & Johannesson, 2009) and thereby enhancing efficiency than do “weak” or no promises. This work raises two questions that motivate the present study.

First, do individuals have intrinsic preferences over language? Are individuals more averse to lies made with strong language than those made with weak language, separate from any effects of the communication on other people? The communication of a truth or a lie can affect the individual making the communication even when there is no discernible audience and when no other person is affected. Several recent papers show that many individuals

¹Recent literature studies drivers and effects of lying behavior. For example (and we apologize for omission of many excellent contributions), scholars have found that the willingness to lie is affected by monetary consequences of the communication (Gneezy, 2005; Gibson et al., 2013), gender (Dreber & Johannesson, 2008; Friesen & Gangadharan, 2012), the extent of the lie (Lundquist, Ellingsen, Gribbe & Johannesson, 2009; Fischbacher & Heusi, 2013), strategic considerations (Sutter, 2009), guilt aversion (Battigalli, Charness & Dufwenberg, 2013; Charness & Dufwenberg, 2010), social cues on how often others lie (Innes & Mitra, 2013), and cooperation in prior play (Ellingsen, Johannesson, Lilja & Zetterqvist, 2009). Effects of lies include less trust (Gawn and Innes, 2017) and more punishment (Brandts and Charness, 2003; Sanchez-Page and Vorsatz, 2007). See the recent survey by Rosenbaum et al. (2014).

are averse to lies in these one-sided situations (see, for example, Lopez-Perez & Spiegelman, 2013; Gibson, Tanner & Wagner, 2013; Fischbacher & Foelmi-Heusi, 2013; Abeler, Becker & Falk, 2014). As argued in much of this work (see, in particular, Mazar, Amir & Ariely, 2008; Gibson et al., 2013), an intrinsic aversion to untruthful communication may stem from values that underpin individual self-concept, the maintenance of which depends upon ethical choices of this type. Our question is: Does the intrinsic preference for honesty depend upon the language used – the strength of the message that defines a lie or truth?

Second, if so – that is, if strong vs. weak language directly affects the preference of the speaker – why might such preferences be advantageous? Extant evidence suggests an advantage of strong promises / strong language in promoting trust and coordination (e.g., CD; Houser & Xiao, 2011; Lundquist et al., 2009). The results beg the question: Why then are promises not *all* “strong”? Why might it be advantageous also to have a weaker aversion to weaker lies (vs. a strong aversion to all lies, for example)?

In this paper, we study how the nature of a communication – the strength of the message that defines a lie or truth – affects subjects’ decisions to lie or not when a lie increases their own payment but harms no one. We find that an exogenous variation in language – from a weak message that states a fact, to a strong message that “solemnly swears” to the fact – has an enormous impact on propensities for truthfulness. Our subjects are much more averse to strong lies than they are to weak lies, an effect that is robust to variations in the monetary reward for dishonesty. The weaker aversion to weaker lies means that weak language promotes our experiment’s socially advantageous lies. We also find that subjects actively prefer to avoid strong language (in favor of weak) when choosing how to lie. The preference over language in turn appears to be related to individuals’ self assessment of their values (Gibson et al., 2013). Higher “moral values” favor stronger language when truth is not at issue, and weaker language when a lie is made.

To our knowledge, this is the first set of experiments to study (1) the effect of language treatments on preferences for honesty, absent any monetary consequences to others, (2) the choice between messages of different strength in communicating a truth or a lie in a

non-strategic setting, and (3) the potential economic advantage of a weaker aversion to weaker lies. Combined with understood benefits of deterring lies in other settings, the third conclusion suggests an evolutionary advantage of preferences that distinguish between the strength of language – a stronger aversion to stronger lies and a weaker aversion to weaker lies. With such preference variation, weak language can be used to promote advantageous lies (as in our first experiment) and strong language can be used to deter lies that are harmful because (for example) they deter trust.

Most closely related to our paper are prior studies on how language variation affects lying/promising behavior in two-sided environments. In these studies, a Receiver takes an action and is affected by a communication that is transmitted by a Sender. A benchmark in this literature is Lundquist, Ellingsen, Gribbe & Johannesson (2009), who compare the effects of weak messages, strong messages and free-form messages (written by Senders) on lies in a two-sided investment game. Our definitions of weak and strong statements closely follow the Lundquist et al. (2009) language. They find that free-form messages significantly reduce the extent of lying relative to either weak or strong messages, and that strong (vs. weak) messages reduce lying for a subset of their sample (students not attending the Stockholm School of Economics). In a trust game, Charness & Dufwenberg (CD) (2006, 2010) find that free-form promises (arguably akin to strong messages) promote more trustworthiness and trust than do “bare promises” (akin to weak messages). Houser & Xiao (2011) use a natural language experiment to distinguish between stronger and weaker promises, finding that only the stronger promises promote trust in CD’s data.² Ellingsen & Johannesson (2004) identify strong effects of free-form promises, while Holm & Danielsson (2005) find insignificant impacts of “bare promises”, also in two-sided trust-type interactions.³ A recent paper by Chen and Houser (2017) studies a modified trust game, focusing on how several features of messages (the use of we/us, reference to money, and length of message) affect the perception of a promise and the likelihood that the promise will be trusted and will be kept; for the three-person game studied, this paper identifies fairly subtle features of

²See Xiao & Houser (2005) for development of the natural language design.

³See also Cappelen, Sorenson & Tungodden (2013) who find that personalizing messages reduces the propensity for deception in an Erat & Gneezy (2012) “white lie” Sender-Receiver game.

messages that make them “stronger” in the eyes of the audience.

When communication occurs in a two-person interaction with consequences for both – as in prior economics studies on language variation (e.g., CD, Lundquist et al., 2009) – many aspects of preferences and behavior come into play. In a trust game, for example, CD argue that effects of communication can be explained by guilt averse preferences, combined with rational second order beliefs about the meaning of promises. The audience (truster) believes that a promise is likely to be honored, increasing her expected return to trust and her disappointment from a violation of trust. The speaker (trustee) in turn is more likely to be trustworthy, by honoring the promise, in order to avoid the guilt aversion cost of disappointing the audience (truster). This logic can explain why a strong message – by producing a rational belief that a promise is more likely to be honored – induces greater truthfulness (promise-honoring) than a weak message, even absent any direct link between preferences and language.

In contrast to the prior literature – and in order to identify effects of language variation on intrinsic preferences for honesty – we focus on one-sided deception experiments devoid of interpersonal and strategic considerations.

2 Experiment 1 (Between Subject)

We implement a variation of the Lopez-Perez & Spiegelman (2013) dot experiment. In the experiment, subjects are exposed to a colored dot at the top of a questionnaire, either blue or green. They are asked to report the color to someone who cannot see the dot and does not know the color. The reports take one of two forms, WEAK or STRONG. With the WEAK form, subjects choose whether to report that “the dot is blue” or that “the dot is green.” With the STRONG form, subjects choose whether to report “I solemnly swear that the dot is” blue or green. The actual dot color is randomly varied from questionnaire to questionnaire. To check for color-blindness, participants are asked to write down the color of the dot at the start of the experiment.⁴

⁴In this check, all students correctly identify the color, with the exception of one missing answer.

We have a two-by-three (six treatment) between-subject design as described in Table 1. The experiment is conducted in eight different economics classes. Each subject participates in only one treatment and plays only one role (sending a message, for example).⁵ There are two message treatments (weak or strong) and three payoff treatments. Payoffs vary in the amount of money the subject earns by reporting the false color of the dot. Under the first payoff variation (Gain \$1), the subject earns \$1 for a truthful report and \$2 for an untruthful report. Similarly, under the second and third payoff variations (Gain \$2 and Gain \$3), the subjects earn \$1 for a truthful report and \$3 or \$4 (respectively) for an untruthful report.

All subjects also choose between a truthful and untruthful report in a baseline where either report pays \$1, so that there is no gain to dishonesty (Gain \$0). Each participant is paid according to either the baseline report or the report made in the positive gain scenario, each with 50 percent probability.⁶

The payoff gains to dishonesty are in line with prior literature. For example, subjects earn one euro more when they are dishonest (vs. honest) in Lopez-Perez & Spiegelman (2013). The payoff gain to dishonesty varies from 0.3 to 1.2 Swiss Francs (roughly 25 U.S. cents to \$1.05) in Gibson et al. (2013) and averages approximately \$2.20 in Fischbacher & Föllmi-Heusi's (2013) baseline experiment, \$1.25 in Cappelen et al. (2013), and at most \$2.60 in Lundquist et al. (2009).⁷

Two features of the design are particularly important. First, decisions in the experiment are made under complete anonymity. Anonymity is stressed both verbally and in written instructions, with subjects told: "Your decisions are completely anonymous. No one, including us, will ever know that you are the person making the decisions you choose." Subjects' decisions are made inside the questionnaire and are not visible to the experimenter when the

⁵In this respect, our experiment differs slightly from Lopez-Perez & Spiegelman (2013), where subjects each play two roles, sending a message and receiving a message.

⁶In all cases, the baseline (Gain \$0) scenario is presented first. We follow the standard approach to avoiding payoff spillovers by randomizing payoffs between the scenarios. The scenario for which a student is paid (zero or positive gain) is determined by a coin flip made after the experiment is completed. For each student, the flip produces either a Heads, to implement scenario one (zero gain from deception), or a Tails, to implement scenario 2 (positive gain), each with equal probability. The procedure is described to the students before the experiment is started.

⁷The latter two averages account for the probability of the Receiver following the Sender's message.

questionnaire is returned. When finished, questionnaires are returned to a box at the front of the room, with the decision sheet covered over by a “front sheet.” Participants are seated facing the front of the room with ample space between them, as in an exam. Students are also told not to communicate with each other for the duration of the experiment, and are monitored to ensure privacy. Hence, a subject’s decisions are not visible to other students at the time choices are being made.

Students are identified for payment based on a registration number indicated on both their questionnaire and an appended tag. Students are instructed to present their tag in class one week after the experiment is completed. Each cash payment is inserted in an envelope with the registration number written on the front; the envelope is given to the student when he or she presents the corresponding tag. Because the experimenter does not know the payment in each envelope at the time the envelopes are collected by participants, there is no way to infer what decisions any particular individual in the experiment has made, even if the experimenter were so inclined. These procedures are fully described to the student participants at the start of the experiment.

Second, although messages are sent to “someone” - other subjects in different classes, who do not observe the dot - there is no other person who responds to or is affected by the choice of message. The receiving subjects are simply given a fixed amount (one dollar), regardless of the message; are told nothing about the context in which the message is chosen; and have no reason to care about the color of the dot.⁸ On the sending side (our interest), subjects are made aware that their choices have no monetary consequence to anyone other than themselves. At the start of the experiment, subjects are (truthfully) told that, following strict professional norms in Economics, (1) everything indicated in the instructions is implemented exactly as stated and (2) there are no consequences of their decisions (to themselves or anyone else) other than those described.⁹ In addition, to avoid any potential subject concern for

⁸The receiving subjects are told (in the case of a weak blue message, for example): “A student in a different classroom observed the color of a dot and chose the following Message: ‘the dot is blue.’” The subject reads and signs for this statement and is paid one dollar. Receiving subjects do not participate in the sending side of the dot experiment, are in a different semester than the senders, and therefore do not know the context in which the message is chosen.

⁹The written instructions also state: “The experiment will be implemented exactly as described... There are no hidden consequences of your decisions.”

dollar costs to the experimenters or a desire to “please” (see Lopez-Perez & Spiegelman, 2013), we convey verbally to the students that we are interested in their decisions for our research and that they should make the decisions that they would like to make under the indicated circumstances. In sum, following prior literature on pure lie aversion, our simple design avoids concerns about responses or consequences to other people. Subjects may well view the situation as hypothetical, as in Gibson et al. (2013), and this is consistent with the one-sided frame that we seek.

The design contrasts with two-sided games in which there are consequences to Receivers and responses by them (Gneezy, 2005; Erat & Gneezy, 2012; Sutter, 2009; Sanchez-Pages & Vorsatz, 2007; and many others). In two-sided studies, experimental instructions actively convey (and thereby encourage subjects to think about) the consequences for others. Like other one-sided deception experiments, ours focuses attention on subjects’ own benefits and costs, both with explicit instructions on the absence of other consequences and the use of a simple narrative.

In some studies on one-sided deception (e.g., Utikal & Fischbacher, 2014; Fischbacher & Föllmi-Heusi, 2013), the experimenter can only infer the extent of deception from statistical properties of resulting choices and does not observe any one individual’s actual decision. In other one-sided experiments, the experimenter observes whether a choice is truthful or not, but the experimenter does not observe the person making the decision (e.g., Lopez-Perez & Spiegelman, 2013, Gibson et al., 2013, and Friesen & Gangadharan, 2012). The same is true in our case. One advantage of this approach is logistical simplicity. More importantly, Gneezy, Kajackaite & Sobel (2017) and Dufwenberg & Dufwenberg (2017) stress that subjects may adjust their behavior in response to believed inferences about the extent and likelihood of truthfulness. The Gneezy et al. (2017) experiments suggest that these forces are at work in “non-observed” treatments when compared to “observed” treatments like ours.¹⁰

¹⁰See also Abeler et al. (2014) and Utikal & Fischbacher (2014), who find “reverse lying” in non-observed treatments. For example, in the Utikal & Fischbacher (2014) sample of nuns, participants report fewer than would be randomly observed of the financially rewarding die rolls. One possible explanation for this finding, cited by the authors, is a subject’s desire to appear truthful to the experimenter, despite monetary costs of doing so. Such appearance motives are unlikely in our case due to subjects’ anonymity.

The experiment is run in eight economics classes at the University of California, Merced. Participation is purely voluntary and has no bearing on course assessment, both of which are indicated to the students before the experiment begins. Course rosters are used to ensure that no student participates more than once.¹¹ Treatments are randomly assigned by mixing of questionnaires. Between 45 and 50 subjects are obtained for each treatment, for a total of 283 student participants. Beyond experimental responses, we solicit each subject's gender and, using the usual incentive-compatible approach, their belief about the proportion of participants who are truthful in their version of the experiment.¹² Sample instructions are provided in the Appendix.¹³ Participants in the experiment received an average payment of about \$7, including a \$5 reward announced at the end of the experiment.

There are minor variations in the male proportion across the treatments. Overall, 49.1 percent of our sample is male; the WEAK treatments are comprised of 47.4 percent males and the STRONG treatments are comprised of 50.7 percent males; the z-statistic for the difference equals 0.54. Blue and green dots are equally distributed in each treatment, with trivial variation across them.

Within each gain scenario (Gain \$1, \$2, and \$3, respectively), the weak and strong treatments are balanced across classes, with only minor variation in numbers of observations within each class between the two (gain-specific, weak and strong) treatments. However, the experiment is conducted in multiple rounds, adding the higher gain scenarios in the later rounds in order to test for robustness of results to higher gains from deception. As a result, caution should be exercised when comparing average behavior *across* the three gain scenarios.

¹¹In any class with an enrollment overlap, students are asked not to participate if they participated in a prior session. Participant sign-up sheets are then compared, after the experiment, to ensure no overlap. Although the post-experiment sign-up sheets (course rosters) indicate the names of participants, anonymity of individual responses is protected; the experimenter does not know and cannot infer what person is associated with any questionnaire.

¹²Each participant is asked to predict the proportion of subjects choosing truthful messages in 5 percentage point bands (0-5%, 5-10%, etc.) and is paid \$1 if the prediction is correct (within five percentage points of actual percentage, plus or minus).

¹³The on-line Appendix provides the introductory instructions describing to students how they are to be paid, how we ensure complete anonymity, that there should be no communication with other students during the experiment, how to ask questions (by raising a hand), etc..

3 Experiment 1 Results

Figure 1 and Table 2 describe the main results from the six treatments. When there is no gain to dishonesty, subjects overwhelmingly choose the truthful message with no noticeable differences across the treatments.¹⁴ Even with weak statements and larger payoff benefits of lies, statistically significant fractions of subjects are truthful ($z=4.95, 3.10, 3.58$ under Gain \$1, \$2, \$3). However, payoff benefits of dishonesty produce significant reductions in the proportion of truthful messages. For example, in the weak message treatments, the proportion of truthful subjects falls from over 92 percent in the zero gain scenario to less than 35 percent when there is a \$1 gain from deception, and to 17 and 22 percent, respectively, when there is a \$2 and \$3 gain from a lie. These results are broadly consistent with prior work documenting that larger payoff benefits from dishonesty draw more subjects to the untruthful option (for example, Gneezy, 2005, and Gibson et al., 2013).¹⁵

Most important for us, strong (vs. weak) messages produce roughly a 30 percentage point increase in the propensity for truth, with strikingly little difference in effect across the different payoffs. Strong messages raise truthfulness by 31.2 percentage points in the \$1 Gain case, 26.2 percentage points in the \$2 Gain case, and 28.9 percentage points in the \$3 Gain case. On average, the increase amounts to more than a doubling in the rate of truthfulness, from 24.5% to 53.5%.

Table 3 reports regression results for the choice of a truthful message (in the positive gain scenarios), controlling for course effects, gender, and alternate payoffs.¹⁶ The first two models are estimated by ordinary least squares, with robust standard errors, and the third is estimated by probit. Both methods provide consistent estimates of marginal effects and stan-

¹⁴Comparing weak and strong treatments in each of three gain scenarios, the corresponding difference between rates of truthfulness in the baseline (weak vs. strong) are, respectively, 0.018 (\$1 Gain subjects, $z=0.52$), -0.022 (\$2 Gain subjects, $z=-0.42$), and -0.044 (\$3 Gain subjects, $z=-0.67$). Differences in behavior across the treatments for the positive gain scenarios can therefore only be attributed to exogenous treatment variation.

¹⁵An interesting feature of the results is the increase in truthfulness when going from a gain of \$2 from deception to a gain of \$3. While there are possible explanations for this effect (including an interaction between payoffs and the impact of truthfulness on self-concept), the observed increase is not statistically significant in our data.

¹⁶In unreported regressions, we also test for any impact of the blue vs. green dot on deception decisions and treatment effects. As expected, we find none.

dard errors. Our main interest is the impact of the overall *strong* treatment indicator. The first model provides the most complete set of other controls, including indicators that allow for different *strong* treatment effects across the alternative payoffs and an interaction between gender and the *strong* message. The estimates reveal no significant difference between the effects of the *strong* message across the different payoffs (with coefficients on *Strong*Gain\$2* and *Strong*Gain\$3* statistically insignificant), and no significant gender interaction effect. In view of these results, the second and third models are more parsimonious specifications.

In all models, the estimates indicate a significant effect of the *strong* (vs. *weak*) treatment. The stronger messages are estimated to raise the rate of truthfulness by between 29.5 and 34.9 percentage points, with $p < 0.01$ in all cases. Larger gains from deception – \$2 vs. \$1, as captured by the *Gain \$2* indicator, and \$3 vs. \$1, as captured by the *Gain \$3* indicator – are estimated to reduce the propensity for truthfulness, consistent with prior work. The male gender is also estimated to reduce truthfulness, as in a number of other studies (e.g., Dreber & Johannesson, 2008; Friesen & Gangadharan, 2012), but not to a significant extent.

Table 4 describes mean beliefs of subjects about the propensity for truthfulness in the different treatments. Although the *strong* treatments engender higher beliefs about truthfulness, the magnitude of the treatment effects on beliefs is quite small. Overall average beliefs of subjects in the *weak* treatments peg the propensity for truthfulness at 36.8 percent, compared with the actual propensity of 24.5 percent in the experiment. Conversely, overall average beliefs in the *strong* treatments predict truthfulness of 45.4 percent, compared with an actual propensity of 53.5 percent in the experiment. While the subjects' predictions are significantly different across the Message treatments (at a two-tail five percent level, $z=2.53$), the mean difference (8.6 percent) is quite small compared with differences in actual propensities (29 percent).

Subjects over-predict in the *weak* treatments (where truthfulness is low) and under-predict in the *strong* treatments (where truthfulness is high). The same is true even when subjects are broken down between truth-tellers and liars;¹⁷ however, the extent of over-

¹⁷We thank an anonymous referee for suggesting this breakdown.

prediction in the weak treatments is smaller for the liars and the extent of under-prediction in the strong treatments is quite small for the truth-tellers. In view of possible “false consensus” effects, that is, beliefs that others act like oneself (Ross et al., 1977), these observations are even more striking and suggest that our behavioral results are unlikely to be driven by differences in perceived norms of conduct across the treatments.¹⁸

4 Experiment 2 (Within Subject)

While the first experiment examines how exogenous variation in message strength affects truthfulness, the second examines choices between messages of different strength.¹⁹ We are interested in whether subjects, when they lie, exhibit an aversion to the choice of a stronger vs. weaker message. If lie aversion is greater when the lie / message is stronger, as indicated in Experiment 1, we should find evidence that (1) lying subjects tend to prefer a weaker statement to a stronger one when payoffs are the same for both, and (2) some lying subjects will be willing to pay a price to substitute a weaker for a stronger lie. To measure the first effect, we must control for any generic preference over language – for example, a preference for a weaker statement when there is no truth-content or payoff consequence to the choice.

4.1 The Experiment

Each participant is shown a dot that is either BLUE or GREEN. Each chooses a message to send to “someone who does not see the dot” and for whom there are no monetary consequences. In each of three scenarios (all three presented to each subject), the participant chooses to send ONE message out of four (4) possibilities. The four messages vary on two dimensions, truth vs. untruth and strong vs. weak:

Message 1. The dot is BLUE.

Message 2. I solemnly swear the dot is BLUE.

¹⁸The beliefs data provide evidence of a false consensus in that predicted rates of truthfulness are significantly higher for truth-tellers than for liars (with z-statistics for the differences equal to 2.77 and 2.06 in the weak and strong pooled treatments, respectively).

¹⁹We are grateful to the referees for suggesting this second experiment.

Message 3. The dot is GREEN.

Message 4. I solemnly swear the dot is GREEN.

The three scenarios vary payoffs associated with the four messages:

Scenario 1 (S1). The payoff is \$1 for any of the four messages.

Scenario 2 (S2). The payoff is \$1 for either of the two truthful messages (Messages 1 and 2 if the color is BLUE, for example) and \$3 for either of the two untruthful messages (Messages 3 and 4 if the color is BLUE, for example).

Scenario 3 (S3). The payoff is \$1 for either of the two truthful messages, \$3 for a weak untruthful message (Message 3, for example) and \$4 for a strong untruthful message (Message 4, for example).

Subjects are paid for one of the three scenarios, each chosen with equal probability. True dot colors, message order, and scenario labels/order are varied across subjects. Procedural protocols are the same as described for Experiment 1.

At the end of the experiment, we ask subjects a set of five questions to provide an indicator of their “protected values” (Gibson et al., 2013). On a one to five scale, they are asked to indicate their agreement or disagreement with each of five statements reflecting moral impulses (where 1 indicates “strongly agree,” 5 indicates “strongly disagree,” and 3 indicates “in between”). A moral index is constructed by summing the rankings from the “moral strength” statements and subtracting the rankings from the “moral flexibility” statements.²⁰ We explore whether the moral index / protected values measure explains message choice.

4.2 Hypotheses

In Scenario 1 (where there is no benefit to lying), we expect and find that almost all subjects are truthful. For the truth tellers, our first (baseline) hypothesis is:

²⁰We consider other measures of morality, based on the survey rankings, but find that the simple moral index has the most explanatory power.

Hypothesis 1 (H1). There is no generic preference for weak vs. strong messages.

We test for H1 using results from Scenario 1 (S1). The associated null hypothesis is that a truth-telling subject chooses a weak vs. strong message in S1 with 50 percent probability (a coin flip).

Our second hypothesis concerns preferences over language among liars:

Hypothesis 2 (H2). There is a tendency for liars to prefer weak (vs. strong) messages when payoffs are unaffected by the chosen language.

H2 reflects a stronger aversion to strong vs. weak lies, as suggested by the results from Experiment 1. There are two tests for H2. First, in Scenario 2 – where any lie produces a net \$2 benefit, regardless of the language – is the proportion of lying participants choosing a weak message significantly higher than 50 percent? Second, we want to control for any generic preference for message strength. For the sample of Scenario 2 lying subjects, is there a significant difference between (1) the preference for a weak vs. strong message in Scenario 2 and (2) the preference for a weak vs. strong message in S1, when there is no truth-content to the choice. ²¹

A third hypothesis identifies a willingness-to-pay to lie with a weak vs. strong message:

Hypothesis 3 (H3). Some lying subjects will prefer to lie with weak (vs. strong) language even when they face a monetary penalty from doing so.

In Scenario 3 (S3), lying with weak (vs. strong) language incurs a \$1 cost. We are interested in whether a significant fraction of subjects nonetheless choose the weak lie, reflecting a positive willingness to pay for the weaker language.

Finally, an individual's protected values / morals – as measured by the elicited moral index values – are expected to affect both truthfulness and preferences over language. Consistent with Gibson et al. (2013), we expect individuals with a higher moral index to be

²¹We make this comparison both for the full set of S2 liars and for the sample of subjects who (1) lie in S2 and (2) tell the truth in S1.

more truthful. We might also expect higher moral index values to produce a greater aversion to stronger lies, and potentially a greater affinity for stronger truths:

Hypothesis 4 (H4). An individual with a higher moral index is more likely to be truthful, to avoid strong lies, and to prefer strong truths.

4.3 Implementation and Results

Experiment 2 is conducted in three economics classes at U.C. Merced using the same general protocols as described for Experiment 1. 52 subjects are enrolled in the experiment, and 59.6 percent of the subject pool is male. Table 5 provides a summary of participant choices between the four message options in the three scenarios. Table 6 provides percentages and test statistics related to our Hypotheses H1-H3. Table 7 presents regressions designed to measure effects of the moral index on message choices.

As expected, almost all subjects (48 of 52) are truthful in the baseline Scenario 1 (S1), where any message decision produces a \$1 payoff. Among the truthful subjects in S1, exactly 50 percent choose a weak (strong) message, consistent with Hypothesis 1. However, as indicated in Table 7, the choice between weak vs. strong messages is affected by individual values as measured by the moral index. Consistent with Hypothesis 4, higher moral index values are associated with a higher likelihood of choosing a Strong message in the baseline (when the message is almost always truthful).

In Scenario 2 (S2), where we add a \$2 benefit to a lie of either strength, 69.2 percent choose to lie and 72.2 percent of the liars choose the weak message. The S2 liars' propensity to choose a weak (vs. strong) message is significantly different from 50 percent ($z=2.98$) and significantly different from their revealed preference for a weak (vs. strong) message in the baseline Scenario 1 ($z=2.092$).²² These results are consistent with our Hypothesis 2 (and results from Experiment 1), indicating a prevailing aversion to strong vs. weak lies.

Scenario 3 (S3) adds a \$1 extra incentive to choose a strong vs. weak lie. The added

²²Comparing S1 and S2 choices for subjects who are both S2 liars and S1 truth tellers yields a similar statistic, $z=2.10$ (see Table 6).

incentive leads to both an increase in dishonest choices (from 69.2 percent in S2 to 76.9 in S3)²³ and a large shift toward the higher-paying strong lie (from 27.8 percent of liars in S2 to 77.5 percent of liars in S3). However, a statistically significant proportion of S3 liars (22.5 percent, $z=3.41$) opt for the weak lie in spite of the associated \$1 cost of avoiding the strong counterpart. Consistent with Hypothesis 3, we thus find that a significant fraction of subjects are willing to pay a positive price to avoid the higher moral cost of a strong vs. weak lie.

Table 7 explores the effect of the self-indicated “protected values” (Gibson et al., 2013) / moral index measure on message choices. Consistent with results of Gibson et al. (2013) and Hypothesis 4, we find a significant positive effect of the moral index on truthfulness in both S2 and S3, as well as the choice of a strong message in the baseline S1. In Scenario 2, we do not find a significant effect of the morality measure on the choice between weak and strong messages for the liars, although the point estimate suggests a negative relationship between individual moral inclinations and the use of strong language to convey a lie. Indeed, controlling for generic preferences over language (from S1), higher moral index values have a significant effect in promoting weak lies in S2 (model (5), $z=2.25$).

In Scenario 3, we expect higher moral impulses to favor weaker lies (among the liars). In S3, the female subjects opting to lie also choose the monetarily advantageous strong lie in almost all cases (16 of 17 female participants); as a result, variation in language choice is only found in the male-only subject pool. For the latter (male S3 liar) sample, we find a significant negative effect of the moral index on the choice of a strong vs. weak lie ($z=-3.04$); that is, a higher moral index is associated with a higher propensity to opt, at a \$1 cost, for the weak lie. Moreover, controlling for language preferences in the baseline S1, the moral index has a significant positive effect in promoting weak lies in the full sample of S3 liars (model (7), $z=2.13$).

Overall, these results reinforce the conclusions from Experiment 1, indicating a greater aversion to stronger vs. weaker lies in a significant fraction of our subject pools. They also

²³The 7.7 percent increase in dishonesty (in S3 vs. S2) is statistically significant at 5 percent ($z=2.06$).

highlight the importance of protected / individual value systems in driving truthfulness and, more to the point for this study, the choice of stronger vs. weaker language in conveying a truth or untruth.

5 Conclusion

In a large body of economics research, communication has been shown to promote economically advantageous coordination and trust. Recent literature documents a primitive attribute of many individuals' preferences that is at least one of the key forces underpinning this phenomenon, namely, an intrinsic aversion to lies. In this paper, we study one-sided deception experiments with no one at the receiving end who is affected by a Sender's message. We find that the nature of a communication – whether it is a strong or weak statement – has a large impact on many subjects' intrinsic preference for honesty.

The implication is that strong (vs. weak) forms of communication can have a much larger effect in promoting truthful exchanges of information. Conversely, weak language can promote socially advantageous lies that benefit someone and harm no one. In this way, lie averse preferences that vary with language, as identified here, potentially permit an advantageous discrimination between situations where lies are harmful (because they impede trust and coordination) and where they are beneficial (as in our experiments).

Our main motivation is to determine whether primitive preferences are directly tied to the language of lies. In two-sided interactions with consequences to both sides, language can affect outcomes due to equilibrium relationships, even absent any direct effects on preferences. Knowing that there is in fact a direct effect (our objective) is relevant to understanding one-sided situations in which only primitive preferences matter. For example, a one-sided setting may be a reasonable representation of circumstances when the “Receiver” of dishonesty is a bureaucracy or corporation that is distant and impersonal (Friesen & Gangadharan, 2012), as in cases of tax reporting, consumer claims on product returns, claims on loan or benefit applications, or self-reports of regulatory infractions.

A link between language and preferences may also be relevant to understanding how

individuals choose language and how language can be designed to promote more effective communication. For example, if there is a stronger aversion to stronger lies (as found here), then stronger language may be chosen in order to persuade others of a truth that is otherwise unlikely to be believed. ²⁴ Such conjectures, while hopefully fueled to some extent by our simple findings, merit more study.

²⁴See, for example, the recent survey by Ozdogan (2016) on the theory of persuasion; Rubinstein (2000), Lipman (2003), and Blume (2000) for fundamental work on language; and DellaVigna and Gentzkow (2010) on experimental evidence.

6 References

- Abeler, J., Becker, A., Falk, A., 2014. Representative evidence on lying costs. *Journal of Public Economics*, 113, 96-104.
- Battigalli, P., Charness, G., Dufwenberg, M., 2013. Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93, 227-233.
- Blume, A., 2000. Coordination and learning with a partial language. *Journal of Economic Theory*, 95, 1-36.
- Brandts, J., Charness, G., 2003. Truth or consequences: An experiment. *Management Science*, 49, 116-130.
- Cappelen, A., Sorenson, E., Tungodden, B., 2013. When do we lie? *Journal of Economic Behavior & Organization*, 93, 258-265.
- Charness, G., Dufwenberg, M., 2006. Promises and partnership. *Econometrica*, 74(6), 1579-1601.
- Charness, G., Dufwenberg, M., 2010. Bare promises: An experiment. *Economic Letters*, 107, 281-283.
- Chen, J., Houser, D., 2017. Promises or lies: Can observers detect deception in written messages. *Experimental Economics*, forthcoming.
- DellaVigna, S., M. Gentzkow, M., 2010. Persuasion: Empirical evidence. *Annual Review of Economics*, 2, 643-669.
- Dreber, A., Johannesson, M., 2008. Gender differences in deception. *Economics Letters*, 99, 197-199.
- Dufwenberg, M., Dufwenberg, M., 2017. Lies in disguise: A theoretical analysis of cheating. Working Paper, University of Arizona.
- Ellingsen, T., Johannesson, M., 2004. Promises, threats and fairness. *Economic Journal*, 114, 397-420.

- Erat, S., Gneezy, U., 2012. White lies. *Management Science*, 58 (4), 723-733.
- Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise - an experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525-547.
- Friesen, L., Gangadharan, L., 2012. Individual level evidence of dishonesty and the gender effect. *Economics Letters*, 117, 624-626.
- Gawn, G., Innes, R., 2017. Do lies erode trust? *International Economic Review*, forthcoming.
- Gibson, R., Tanner, C., Wagner, A.F., 2013. Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review*, 103(1), 532-48.
- Gneezy, U. 2005. Deception: The role of consequences. *American Economic Review*, 95, 384-94.
- Gneezy, U., Kajackaite, A., Sobel, J., 2017. Lying aversion and the size of the lie. Working Paper, U.C.S.D. and U.S.C..
- Houser, D., Xiao, E., 2011. Classification of natural language messages using a coordination game. *Experimental Economics*, 14, 1-14.
- Holm, H., Danielsson, A., 2005. Fairness and promises for sale. Working paper, Lund University, 2005.
- Innes, R., Mitra, A., 2013. Is dishonesty contagious? *Economic Inquiry*, 51(1), 722-734.
- Lipman, B., 2003. Language and economics. In: *Cognitive Processes and Economic Behavior*, Ed.s: N. Dimitri, M. Basili, & I. Gilboa, 75-93.
- Lopez-Perez, R., Spiegelman, E., 2013. Why do people tell the truth? Experimental evidence for pure lie aversion. *Experimental Economics*, 16(3), 233-247.
- Lundquist, T., Ellingsen, T., Gribbe, E., Johannesson, M., 2009. The aversion to lying. *Journal of Economic Behavior & Organization*, 70 (1-2), 81-92.

- Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45, 633-44.
- Matlock, T., 2012. Framing political messages with grammar and metaphor. *American Scientist*, 100, 478-83.
- Ozdogan, A., 2016. A survey of strategic communication and persuasion. *Bogazici Journal Review of Social, Economic & Administrative Studies*, 30, 1-21.
- Rosenbaum, S., Billinger, S., Stieglitz, N., 2014. Lets be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45, 181-96.
- Ross, L., Greene, D., House, P., 1977. The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13, 279-301.
- Rubinstein, A., 2000. *Economics and Language*, Cambridge: Cambridge University Press.
- Sánchez-Pagés, S., Vorsatz, M., 2007. An experimental study of truth-telling in a sender-receiver game. *Games & Economic Behavior*, 61, 86-112.
- Sutter, M., 2009. Deception through telling the truth? Experimental evidence from individuals and teams. *Economic Journal*, 119, 4760.
- Talmy, L., 2000. *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.
- Utikal, V., Fishbacher, U., 2013. Disadvantageous lies in individual decisions. *Journal of Economic Behavior & Organization*, 85, 108-111.
- Xiao, E., Houser, D., 2005. Emotion expression in human punishment behavior. *Proceedings of the National Academy of Science*, 102, 7398-7401.

Figure 1. Experiment 1 Results

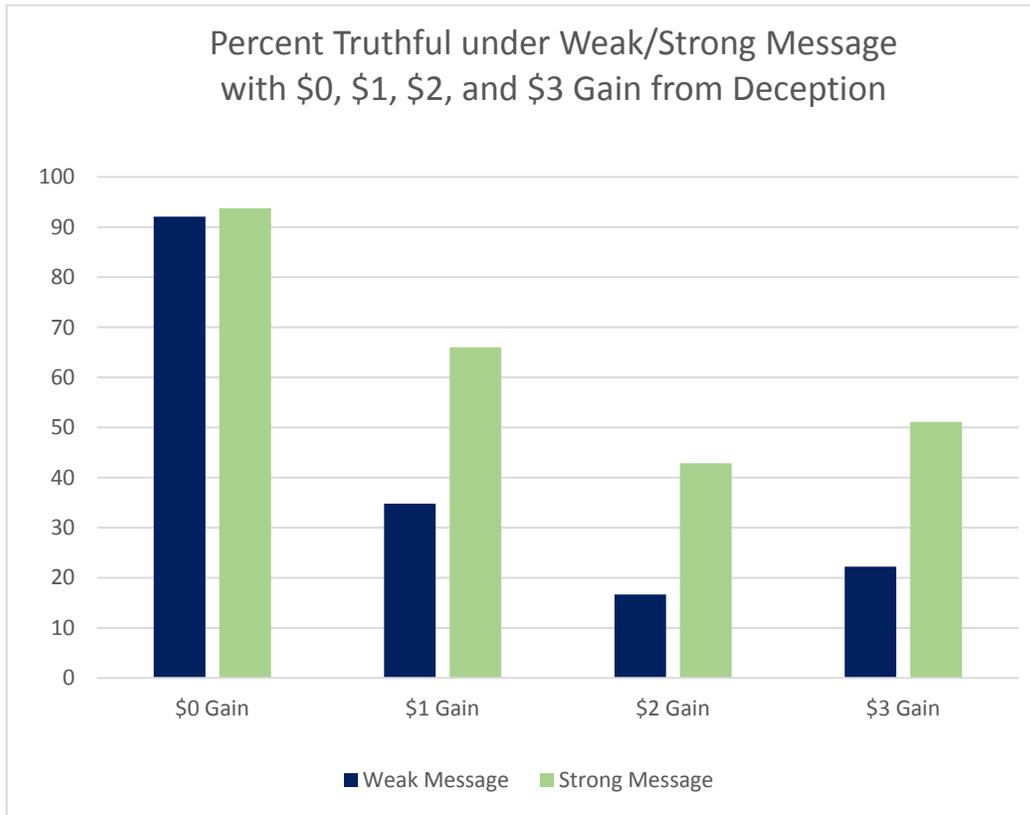


Figure 2. Experiment 2 Results

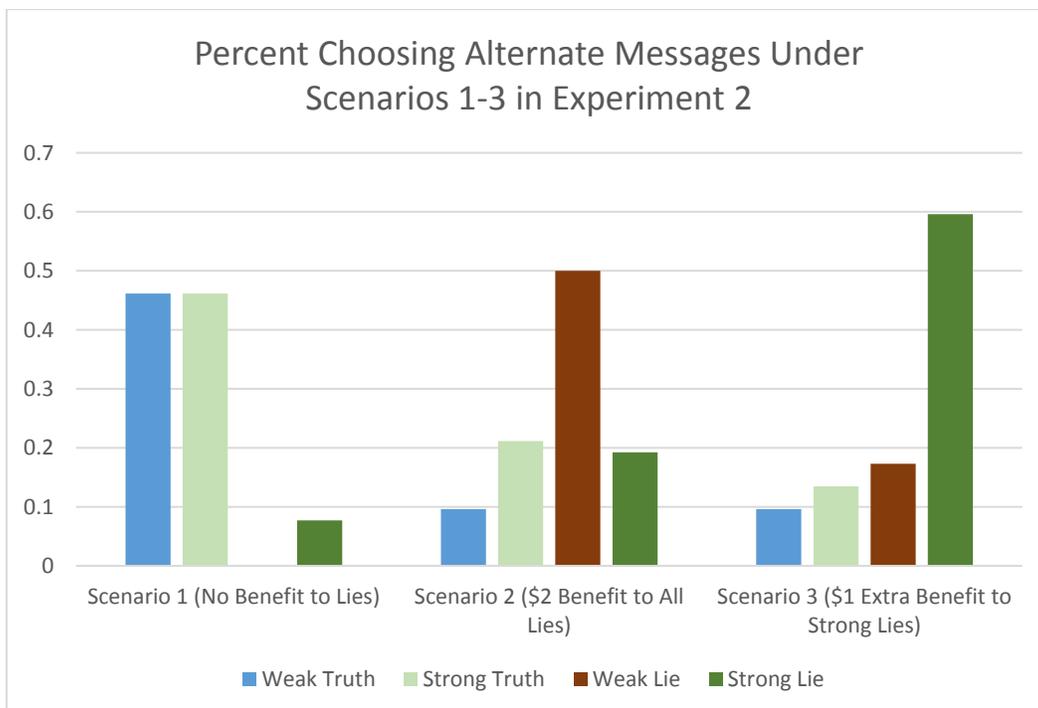


Table 1. Treatments in Experiment 1

Payoffs from (Truthful Message, Untruthful Message)

Message	Gain from Deception		
	Gain \$1	Gain \$2	Gain \$3
↓			
Weak	(\$1,\$2)	(\$1,\$3)	(\$1,\$4)
Strong	(\$1,\$2)	(\$1,\$3)	(\$1,\$4)

Table 2. Experiment 1 Results

Treatment					
Message	Gain from Deception	N	Percent Truthful	Difference: Strong - Weak	z-statistic
Weak	\$1	46	34.78		
Strong	\$1	50	66.00	31.22	3.22***
Weak	\$2	48	16.67		
Strong	\$2	49	42.86	26.19	2.95***
Weak	\$3	45	22.22		
Strong	\$3	45	51.11	28.89	2.98***
Weak	\$1-\$3	139	24.46		
Strong	\$1-\$3	144	53.47	29.01	5.25***
Weak	\$0	139	92.09		
Strong	\$0	144	93.75	1.66	0.54

Notes: *** p<0.01 (two-tail). Weak Message: "The dot is blue/green."

Strong Message: "I solemnly swear that the dot is blue/green."

Table 3. Probit Regressions for Truthful Message, Experiment 1

	(1)	(3)
	Marg Effect	Marg Effect
Strong	0.3427 (3.02)***	0.3043 (5.12)***
Strong * Gain \$2	-0.0259 (-0.18)	--
Strong * Gain \$3	-0.0181 (-0.12)	--
Gain \$2	-0.1538 (-1.31)	-0.2138 (-3.06)***
Gain \$3	-0.0659 (-0.45)	-0.1408 (-1.86)*
Male	-0.0194 (-0.21)	-0.0582 (-0.95)
Male * Strong	-0.0534 (-0.44)	--
Course Effects	Yes	Yes

N=281. Dependent Variable: Truthful Message (1=Truth, 0=Lie). Robust t-statistics in parentheses. Marg. effects at means. *p<0.10, *** p<0.01.

Table 4. Subject Beliefs About the Propensity for Truth, Experiment 1

Treatment						
Message	Gain from Deception	N	Mean (Percent)	Standard Deviation	Difference: Strong-Weak	z-statistic
Weak	\$1	46	35.76	28.27		
Strong	\$1	50	45.30	28.01	9.54	1.66
Weak	\$2	48	34.27	28.25		
Strong	\$2	49	42.19	27.09	7.92	1.41
Weak	\$3	45	40.61	31.16		
Strong	\$3	45	49.05	28.72	8.44	1.34
Weak	\$1-\$3	139	36.82	29.14		
Strong	\$1-\$3	144	45.42	27.87	8.60	2.53**
Weak	\$0	138	74.69	28.23		
Strong	\$0	144	76.08	25.97	1.39	0.43
Truth-tellers						
Weak	\$1-\$3	34	47.65	25.03		
Strong	\$1-\$3	77	49.90	24.25	2.25	0.44
Liars						
Weak	\$1-\$3	105	33.31	26.92		
Strong	\$1-\$3	67	40.26	30.91	6.95	1.46

Beliefs measured at midpoints of chosen percentage bands. ** p<0.05.

Table 5. Experiment 2 Results

Message Choice↓	Scenario 1 (S1)		Scenario 2 (S2)		Scenario 3 (S3)	
	Payoffs	Obs.	Payoffs	Obs.	Payoffs	Obs.
Weak Truth	\$1	24	\$1	5	\$1	5
Strong Truth	\$1	24	\$1	11	\$1	7
Weak Lie	\$1	0	\$3	26	\$3	9
Strong Lie	\$1	4	\$3	10	\$4	31
Total Obs.		52		52		52

Notes: "Obs." indicate the number of subjects choosing each of the four possible messages in each scenario.

Table 6. Test Statistics for Experiment 2 Hypotheses

Sample	Scenario	Hypothesis	Proportion Choosing Weak		z-statistic
S1 Truth Tellers (N=48)	S1	H1	0.5000		0.00 (different from 50%)
S2 Liars (N=36)	S2	H2	0.7222		2.98*** (different from 50%)
S3 Liars (N=40)	S3	H3	0.2250		3.41*** (different from zero)
			S1 Strong	(S1 Strong)-(S2 Strong) ⁺	
			Mean	Mean Diff. (Std. Dev.)	z-statistic (Difference=0)
S2 Liars (N=36)	S1,S2	H2	0.5000	0.2222 (0.6374)	2.092**
S2 Liars and S1 Truth Tellers (N=32)	S1,S2	H2	0.4375	0.2500 (0.6720)	2.104**

** , *** denotes significant (two-tail) at 5%, 1%. ⁺ (S1 Strong)-(S2 Strong) = difference between each subjects' choice of a strong (=1) vs. weak (=0) message in S1 and the corresponding strong (=1) vs. weak (=0) choice in S2. The sample standard deviation for the difference is indicated in parenthesis.

Table 7. Experiment 2 Regressions: Effects of the Moral Index

	(1)	(2)	(3)	(4)	(5)	(6)	(6) ⁺
Dep. Var. →	S1 Strong	S2 Truth	S3 Truth	S2 Strong	S1 Strong - S2 Strong	S3 Strong	S1 Strong - S3 Strong
Estimation →	Probit				OLS	Probit	OLS
Sample →	All	All	All	S2 Liars	S2 Liars	Male S3 Liars ⁺	S3 Liars
N →	51	51	51	35	35	23	39
	Marg. Eff (z-stat) ↓	Marg. Eff (z-stat) ↓	Marg. Eff (z-stat) ↓				
Male	-0.053 (-0.35)	0.100 (0.73)	0.110 (0.86)	-0.252 (-1.55)	0.1744 (0.77)	----	0.3271 (1.50)
Moral Index	0.050 (2.55)**	0.038 (1.96)*	0.040 (2.39)**	-0.021 (-1.08)	0.0575 (2.25)**	-0.116 (-3.039)***	0.0658 (2.13)**
Course Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust t-statistics in parentheses. Marginal effects at means. Moral Index has sample mean (standard deviation) 4.59 (3.53), varying from -4 to 13. *p<0.10, **p<0.05, ***p<0.01.

⁺ In S3, 16 of 17 female liars choose Strong messages. Model (6) is limited to male S3 liars who exhibit variation in message choice.

Sample Instructions for Experiment 2 (Blue dot)

In this short experiment you will be paid for ONE of three SCENARIOS. The scenario for which you will be paid will be chosen at random.

There is a colored dot at the top of this page. What color is the dot? _____

Here are four Messages that you can send about the dot to someone who cannot see the dot and does not know its color (and for whom there are no monetary consequences).

Message 1: The dot is BLUE.

Message 2: I solemnly swear that the dot is BLUE.

Message 3: The dot is GREEN.

Message 4: I solemnly swear that the dot is GREEN.

Your choice of Message determines your PAYMENT in each of the following 3 scenarios. In each case, the payment is listed beside the corresponding Message.

For each of the following scenarios, please circle the Message you choose to send.

SCENARIO 1: You will get **\$1** if you send **ANY** of the 4 Messages. ****I choose to send (please CIRCLE ONE)****

Message 1 (Payment \$1)

(The dot is BLUE)

Message 2 (Payment \$1)

(I solemnly swear that the dot is BLUE)

Message 3 (Payment \$1)

(The dot is GREEN)

Message 4 (Payment \$1)

(I solemnly swear that the dot is GREEN)

SCENARIO 2: You will get **\$1** if you send **Message 1 OR 2 (BLUE dot)** and **\$3** if you send **Message 3 OR 4 (GREEN dot)**. ****I choose to send (please CIRCLE ONE)****

Message 1 (Payment \$1)

(The dot is BLUE)

Message 2 (Payment \$1)

(I solemnly swear that the dot is BLUE)

Message 3 (Payment \$3)

(The dot is GREEN)

Message 4 (Payment \$3)

(I solemnly swear that the dot is GREEN)

SCENARIO 3: You will get **\$1** if you send **Message 1 OR 2 (BLUE dot)**, **\$3** if you send **Message 3 (GREEN dot)**, and **\$4** if you send **Message 4 (GREEN dot)**. ****I choose to send (please CIRCLE ONE)****

Message 1 (Payment \$1)

(The dot is BLUE)

Message 2 (Payment \$1)

(I solemnly swear that the dot is BLUE)

Message 3 (Payment \$3)

(The dot is GREEN)

Message 4 (Payment \$4)

(I solemnly swear that the dot is GREEN)

Please indicate the extent to which you agree or disagree with the following statements:

1) Cheating is something

a) about which one should be flexible if the situation demands (PLEASE CIRCLE ONE)

STRONGLY AGREE	AGREE	IN BETWEEN	DISAGREE	STRONGLY DISAGREE
-------------------	-------	------------	----------	----------------------

b) that should not be done, no matter the benefit (PLEASE CIRCLE ONE)

STRONGLY AGREE	AGREE	IN BETWEEN	DISAGREE	STRONGLY DISAGREE
-------------------	-------	------------	----------	----------------------

2) Being a moral person is very important to me (PLEASE CIRCLE ONE)

STRONGLY AGREE	AGREE	IN BETWEEN	DISAGREE	STRONGLY DISAGREE
-------------------	-------	------------	----------	----------------------

3) I respect people who are pragmatic and do what is needed to get something done, even if that means bending a few rules and blurring some moral lines (PLEASE CIRCLE ONE)

STRONGLY AGREE	AGREE	IN BETWEEN	DISAGREE	STRONGLY DISAGREE
-------------------	-------	------------	----------	----------------------

4) Religion is important in my life (PLEASE CIRCLE ONE)

STRONGLY AGREE	AGREE	IN BETWEEN	DISAGREE	STRONGLY DISAGREE
-------------------	-------	------------	----------	----------------------

FINAL QUESTION:

5) Can you tell us your gender (please CIRCLE ONE):

Male

Female