Abeler, Johannes; Nosenzo, Daniele; Raymond, Collin

Working Paper

Preferences for truth-telling

CeDEx Discussion Paper Series, No. 2016-13

Provided in Cooperation with:
The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx)

Suggested Citation: Abeler, Johannes; Nosenzo, Daniele; Raymond, Collin (2016) : Preferences for truth-telling, CeDEx Discussion Paper Series, No. 2016-13, The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx), Nottingham

This Version is available at:
https://hdl.handle.net/10419/163014

CEDEX

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Discussion Paper No. 2016-13

Johannes Abeler, Daniele Nosenzo and Collin Raymond

September 2016

Preferences for truth-telling

CeDEx Discussion Paper Series

ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
Fax: +44 (0) 115 95 14159
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

Preferences for truth-telling

Johannes Abeler, Daniele Nosenzo and Collin Raymond*

September 6, 2016

Abstract

Private information is at the heart of many economic activities. For decades, economists have assumed that individuals are willing to misreport private information if this maximizes their material payoff. We combine data from 72 experimental studies in economics, psychology and sociology, and show that, in fact, people lie surprisingly little. We then formalize a wide range of potential explanations for the observed behavior, identify testable predictions that can distinguish between the models and conduct new experiments to do so. None of the most popular explanations suggested in the literature can explain the data. We show that only combining a preference for being honest with a preference for being seen as honest can organize the empirical evidence.

Keywords: private information, honesty, truth-telling, lying, meta study

JEL Codes: D03, D82, H26, I13, J31

*Abeler: University of Oxford, IZA and CESifo (e-mail: johannes.abeler@economics.ox.ac.uk); Nosenzo: University of Nottingham (e-mail: Daniele.Nosenzo@nottingham.ac.uk); Raymond: Amherst College (e-mail: collinbraymond@gmail.com). JA thanks the ESRC for financial support under grant ES/K001558/1. We thank Steffen Altmann, Steve Burks, Gary Charness, Vince Crawford, Armin Falk, Urs Fischbacher, Simon Gächter, Philipp Gerlach, David Gill, Uri Gneezy, Andreas Grunewald, David Huffman, Navin Kartik, Michael Kosfeld, Erin Krupka, Dmitry Lubensky, Daniel Martin, Takeshi Murooka, Simone Quercia, Heiner Schuhmacher, Daniel Seidman, Klaus Schmidt, Joel Sobel, Marie Claire Villeval and Joachim Winter for helpful discussions. Many valuable comments were also received from numerous seminar and conference participants. We are very grateful to all authors who kindly shared their data for the meta study: Yuval Arbel, Alessandro Bucciol, Christopher Bryan, Julie Chytilová, Sophie Clot, Doru Cojoc, Julian Conrads, Daniel Effron, Anne Foerster, Toke Fosgaard, Simon Gächter, Holger Gerhardt, Andreas Glöckner, Joshua Greene, Benni Hilbig, David Hugh-Jones, Ting Jiang, Elina Khachatryan, Martina Kroher, Alan Lewis, Michel Marechal, Gerd Muehlheusser, David Pascual Ezama, Eyal Pe'er, Marco Piovesan, Matteo Ploner, Wojtek Przepiorka, Heiko Rauhut, Tobias Regner, Rainer Rilke, Andreas Roider, Bradley Ruffle, Anne Schielke, Jonathan Schulz, Shaul Shalvi, Jan Stoop, Bruno J. Verschuere, Berenike Waubert de Puiseau, Niklas Wallmeier, Joachim Winter and Tobias Wolbring. Felix Klimm, Ines Lee, Felix Samy Soliman, David Sturrock, Kelly Twombly and James Wisson provided outstanding research assistance. Ethical approval for the experiments was obtained from the Nottingham School of Economics Research Ethics Committee and the Nuffield Centre for Experimental Social Sciences Ethics Committee.

Reporting private information is at the heart of many economic activities, for example, a self-employed shopkeeper reporting her income to the tax authorities (e.g., Allingham and Sandmo 1972), a doctor stating a diagnosis (e.g., Ma and McGuire 1997), or an expert giving advice (e.g., Crawford and Sobel 1982). For decades, economists have assumed that people are not intrinsically concerned about lying or telling the truth and that their utility only depends on their material payoff. This implies that, if misreporting cannot be detected, individuals should always submit the report that maximizes their material gains.

Until recently, the assumption of always submitting the payoff-maximizing report has gone basically untested, partly because empirically studying reporting behavior is by definition difficult. In the last years, a fast growing experimental literature across economics, psychology and sociology has begun to study patterns of reporting behavior empirically and a string of theoretical papers has been built on the assumption of some preference for truth-telling (e.g., Kartik et al. 2007, Matsushima 2008, Ellingsen and Östling 2010, Kartik et al. 2014b).

In this paper, we aim to deepen our understanding of how people report private information. We first document that behavior in the experimental literature is indeed far from the assumption of payoff-maximizing reporting. Then, in the main contribution of the paper, we formalize a wide range of explanations for this aversion to lying, identify testable predictions to distinguish the models, and test them in new experiments.

In order to cleanly identify the motivations driving aversion to lying, we focus on a setting without strategic interactions, i.e., we abstract from sender-receiver games or verification of messages, such as audits. We therefore use the experimental paradigm introduced by Fischbacher and Föllmi-Heusi (2013) (for related methods, see Batson et al. 1997 and Warner 1965): subjects privately observe the outcome of a random variable, report the outcome and receive a monetary payoff proportional to their report. While no individual report can be identified as truthful or not, the researcher can judge the reports of a group of subjects. This paradigm is the one used most widely in the literature and several recent studies have shown that behavior in it correlates well with cheating behavior outside the lab (Hanna and Wang 2013, Cohn and Maréchal 2015, Cohn et al. 2015, Gächter and Schulz 2016b, Potters and Stoop 2016, Dai et al. forthcoming).[1]

[1]Three other paradigms are also widely used in the literature. In the sender-receiver game, introduced by Gneezy (2005), one subject knows which of two states is true and tells another subject (truthfully or

In the first part of our paper (Section 1 and Appendix A), we combine data from 72 studies that use setups akin to Fischbacher and Föllmi-Heusi (2013), involving more than 32000 subjects across 43 countries. Our study is the first quantitative meta analysis of this experimental paradigm. We show that subjects forgo on average about three-quarters of the potential gains from lying. This is a very strong departure from the standard economic prediction and a much larger deviation than many other widely discussed non-standard behaviors, like altruism or reciprocity.[2] This strong preference for truth-telling is robust to increasing the payoff level 500-fold or repeating the reporting decision up to 50 times.[3] The cross-sectional patterns of reports are extremely similar across studies. Overall, we document a stable and coherent corpus of evidence across many studies, which could potentially be explained by one unifying theory.

In the second part of the paper (Section 2 and Appendices B and C), we formalize a wide range of explanations for the observed behavior, including all of the many explanations that have been suggested, often informally, in the literature. The classes of models we consider cover three broad types of plausible motivations: a direct cost of lying (e.g., Ellingsen and Johannesson 2004, Kartik 2009); a valuation of some kind of reputation linked to the report (e.g., Mazar et al. 2008); and the influence of social norms and social comparisons, including

not) which one it is. The other subject then chooses an action. Payoffs are determined by the state and the action. The advantage is that the experimenter knows the true state and can thus judge individually whether a subject lied or not. The strategic interaction makes the setting more complex though, especially if one is interested in studying the underlying motives of reporting behavior. In the "matrix task", introduced by Mazar et al. (2008) (and similar real-effort reporting tasks, e.g., Ruedy and Schweitzer (2010)), subjects solve a mathematical problem, are then given the correct set of answers and report how many answers they got right. Finally, they destroy their answer sheet, making lying undetectable. This setup is quite similar to Fischbacher and Föllmi-Heusi (2013) but has the advantage of being less abstract. It does add ambiguity about the truthful proportion of correct answers in the population which makes testing theories harder. In Charness and Dufwenberg (2006), subjects can send a message promising (or not) a particular future action. Incorrect messages can thus be identified for each subject ex-post. Charness and Dufwenberg show that the message affects the action, the truthfulness of the message at the time of sending is thus unclear. Other influential experiments in this literature are, e.g., Ellingsen and Johannesson (2004) and Vanberg (2008).

[2] Altruism is often measured by the amount given in dictator-game experiments. There, subjects transfer on average 28.4 percent of the pie, i.e., they forgo only about a quarter of the potential gains from the experiment (Engel 2011). Positive reciprocity is often measured by second-mover behavior in trust games. There, subjects send on average about 38 percent of their endowment back (Johnson and Mislin 2011; Cardenas and Carpenter 2008), forgoing about one third of the possible monetary gains. Negative reciprocity is often measured by second-mover behavior in ultimatum-game experiments. There, the average rejection rate is 15.8 percent. Since lower offers are rejected more readily, the actual gains forgone are lower than this share (Oosterbeek et al. 2004).

[3] In most experiments using this paradigm, the money obtained by reporting comes from the experimenter but there are almost a dozen studies in which the money comes from another subject and behavior is very similar, see Appendix A for details.

guilt aversion (e.g., Weibull and Villa 2005, Charness and Dufwenberg 2006). We also consider numerous extensions, combinations and mixtures of the aforementioned models (e.g., Kajackaite and Gneezy 2015, Boegli et al. 2016) including several new models that, to us, seemed plausible. For all models we make minimal assumptions on the functional form and allow for full heterogeneity of preference parameters, thus allowing us to derive very general conclusions. Importantly, the previous literature cannot distinguish between these models. All of the models are able to explain the central finding that we document in the meta study, namely that many people refrain from lying maximally. Many models are also able to explain several of the other findings of the meta study, e.g., that non-maximal states are reported more often than their true likelihood.

One of the key insights of our paper is identifying five testable predictions that can differentiate between the models. We show that the models differ in (i) how the distribution of true states affects one's report; (ii) how the belief about the reports of other subjects influences one's report[4]; (iii) whether the observability of the true state affects one's report; (iv) whether some subjects will lie downwards, i.e. report a state that yields a lower payoff than their true state, when the true state is observable; (v) whether all reports are made with positive probability. Our predictions come in two varieties: (i) to (iii) are comparative statics while (iv) and (v) concern properties of equilibrium behavior.

We take a Popperian approach in our empirical analysis (Popper 1934). Each of our tests, taken in isolation, is not able to pin down a particular model. However, each test is able to cleanly falsify whole classes of models and all tests together allow us to tightly restrict the set of models that can explain the data. Since we formalize a large number of models (a total of 22), covering a broad range of potential motives, the set of surviving models is more informative than if we had only falsified a single model, e.g., the standard model. The surviving set obviously depends on the set of models and the empirical tests that we consider. However, the transparency of the falsification process allows researchers to easily adjust the set of non-falsified models as new evidence becomes available.

In the third part of the paper (Section 3 and Appendix D), we conduct new laboratory experiments with more than 1600 subjects to test between the classes of models. The ex-

[4]Technically, for some models this test works through updating the belief about the distribution of other subjects' preferences. For other models, it works through directly changing the best response of subjects (see Section 2 for details).

periments focus on the first four predictions outlined above. Regarding prediction (v), the meta study establishes the result that in virtually all treatments in the literature (including our new experiments) all reports are made with positive probability. This finding rules out most of the models in which individuals only care about their reputation of having reported truthfully since these models often predict that all subjects pool on reporting the same state.

To test the influence of the distribution of true states (prediction (i)), we let subjects draw from an urn with two states and we change the probability of drawing the high-payoff state between treatments. Our comparative static is 1 minus the ratio of low-payoff reports to expected low-payoff draws. Under the assumption that individuals never lie downwards, this can be interpreted as the fraction of individuals who lie upwards. We find a very large treatment effect. When we move the share of true high-payoff states from 10 to 60 percent, the share of subjects who lie up increases by almost 30 percentage points. We replicate this result in a second experiment with ten potential states. This result falsifies direct lying-cost models because this cost only depends on the comparison of the report to the true state that was drawn but not on the prior probability of drawing the state.[5]

To test the influence of subjects' beliefs about what others report (prediction (ii)), we use anchoring, i.e., the tendency of people to use salient information to start off one's decision process (Tversky and Kahneman 1974). By asking subjects to read a description of a "potential" experiment and to "imagine" two "possible outcomes" which differ by treatment, we are able to shift (incentivized) beliefs of subjects about the behavior of other subjects by more than 20 percentage points. This change in beliefs does not affect behavior: subjects in the high-belief treatment are slightly less likely to report the high state, but this is far from significant. This result rules out all the social-comparisons models we consider. In these models, individuals prefer their outcome or behavior to be similar to that of others, so if they believe others report the high state more often they want to do so, too.

To test the influence of the observability of the true state (prediction (iii)), we implement the random draw on the computer and are thus able to recover the true state. We use a double-blind procedure to alleviate subjects' concerns about indirect monetary consequences of lying, e.g., being excluded from future experiments. We find significantly less over-reporting

[5] Gneezy, Kajackaite, and Sobel (personal communication) and Garbarino, Slonim, and Villeval (personal communication) have also run experiments in which they vary the distribution of true states. Their findings are consistent with our results.

in the treatment in which the true state is observable compared to when it is not. Moreover, we find that no subject lies downwards in this treatment (prediction (iv)). These findings are again inconsistent with lying-cost models and social-comparison models since in those models utility does not depend on the observability of the true state.

In Section 4, we compare the predictions of the models to the gathered empirical evidence. The main empirical finding is that our five tests rule out almost all of the models previously suggested in the literature. From the set of models we consider, only combining a preference for being honest with a preference for being seen as honest (or a model whose intuition and prediction are very similar) can explain the data. This intuition is also present in Khalmetski and Sliwka (2016) and Frankel and Kartik (2016). We then turn to calibrating a simple version of our model, showing that it can match all the stylized facts we uncovered in the meta study and the patterns in our new experiments. In the model, individuals suffer a cost that is proportional to the monetary gain from lying and a cost that is linear in the probability that they lied (given their report and the equilibrium report). Both cost components are important.

Section 5 concludes and discusses policy implications. Three key insights follow from our study. First, our meta analysis shows that the data are not in line with the assumption of payoff-maximizing reporting but rather with some preference for truth-telling. Second, our results suggest that these preferences are comprised of two components, a direct cost of lying that depends on the comparison between the report and the true state, and a reputational cost for being thought of as a liar. This contrasts with most of the previous literature on lying aversion which focuses on a single cost component. Finally, policy interventions that rely on voluntary truth-telling by some participants could be very successful, in particular if it is made hard to lie while keeping a good reputation.

1 Meta study

1.1 Design

The meta study covers 72 experimental studies containing 362 treatment conditions that fit our inclusion criteria. We include all studies using the setup introduced by Fischbacher and Föllmi-Heusi (2013) (which we will refer to as "FFH paradigm"), i.e., in which subjects

conduct a random draw and then report their outcome of the draw, i.e., their state. We require that the true state is unknown to the experimenter but that the experimenter knows the distribution of the random draw. We also include studies in which subjects report whether their prediction of a random draw was correct (as in Jiang 2013). The payoff from reporting has to be independent of the actions of other subjects, but the reporting action can have an effect on other subjects. The expected payoff level must not be constant, e.g., no hypothetical studies, and subjects are not allowed to self-select into the reporting experiment after learning about the rules of the experiment. For more details on the selection process, see Appendix A.

We contacted the authors of the identified papers and thus obtained the raw data of 46 studies. For the remaining studies, we extract the data from graphs and tables shown in the papers. This process does not allow to recover additional covariates for individual subjects, like age or gender, and we cannot trace repeated decisions by the same subject. However, for most of our analyses, we can reconstruct the relevant raw data entirely in this way. The resulting data set thus contains data for each individual subject. Overall, we collect data on 108140 decisions by 32503 subjects. Experiments were run in 43 countries which cover 68 percent of world population and 81 percent of world GDP. A good half of the overall sample are students, the rest consists of representative samples or specific non-student samples like children, bankers or nuns. Table A.4 lists all included studies. Studies for which we obtained the full raw data are marked by *.

Having access to the (potentially reconstructed) raw data is a major advantage over more standard meta studies. We can treat each subject as an independent observation, clustering over repeated decisions and analyzing the effect of individual-specific co-variates. More importantly, we can separately use within-treatment variation (by controlling for treatment fixed effects), within-study variation (by controlling for study fixed effects) and across-study variation for identification.[6]

Since the potential reports differ widely between studies, e.g., sides of a coin or color of balls drawn from an urn, we focus on the payoff consequences of a report as its defining characteristic. To make the different studies comparable, we map all reports into a "standardized report". Our standardized report has three key properties: (i) if a subject's report leads to

[6]For other meta studies using the full individual subject data (albeit on different topics), see e.g., Harless and Camerer (1994) or Weizsäcker (2010).

the lowest possible payoff, the standardized report is -1, (ii) if the report leads to the highest possible payoff, it is $+1$ and (iii) if the report leads to the same payoff as the expected payoff from truthful reporting, the standardized report is 0. In particular we define:

$$r_{standardized} = \frac{\pi - \pi^{truthful}}{\pi^{truthful} - \pi^{min}} \quad \text{if} \quad \pi < \pi^{truthful}$$
$$r_{standardized} = \frac{\pi - \pi^{truthful}}{\pi^{max} - \pi^{truthful}} \quad \text{if} \quad \pi \geq \pi^{truthful}$$

where π is the payoff of a given report, π^{min} the payoff from reporting the lowest possible state, π^{max} the payoff from reporting the highest state and $\pi^{truthful}$ is the expected payoff from truthful reporting. For example, a roll of a six-sided die would result in standardized reports of -1, -0.6, -0.2, $+0.2$, $+0.6$, or $+1$.[7]

In general, without making further assumptions, one cannot say how many people lied or by how much in the FFH paradigm. We can only say how much money people left on the table. An average standardized report greater than 0 means that subjects leave less money on the table than a group of subjects who report fully honestly.

1.2 Results

Finding 1 *Subjects obtain only about a quarter of the payoff they could obtain by reporting the payoff-maximizing state.*

Figure 1 depicts an overview of the data. Standardized report is on the y-axis and the available monetary payoff, i.e., $\pi^{max} - \pi^{min}$, is on the x-axis (converted by PPP to 2015 USD). As payoff, we take the expected payoff, i.e., the nominal payoff used in the experiment times the probability that a subject receives the payoff, in case not all subjects are paid. Each bubble represents the average standardized report of one treatment. The size of the bubble is proportional to the number of subjects in that treatment. The baseline treatment of Fischbacher and Föllmi-Heusi (2013) is marked in the figure. It replicates quite well.

If all subjects were rational, liked money and had no concerns about lying, all bubbles would be at $+1$. In contrast, we find that the average standardized report is only 0.216. This means that subjects forego about three-quarters of the potential gains from lying. This

[7] For uniform state distributions with linear payoff increases, the standardized report simplifies to a linear transformation of $(\pi - \pi^{min})/(\pi^{max} - \pi^{min})$.

is a very strong departure from the standard economic prediction. It is also a much larger deviation than other widely discussed non-standard behaviors, like altruism or reciprocity.

This finding turns out to be quite robust.

Finding 2 *Subjects continue to refrain from lying maximally when stakes are increased. Indeed, we find almost no increase in the average report when stakes are increased 500-fold.*

Figure 1 shows that an increase in incentives affects behavior only very little. In our sample, the potential payoff from misreporting ranges from cents to 50 USD, a 500-fold increase. The fitted regression line in the figure is from a quadratic regression (the x-axis is on a log scale) and is slightly hump-shaped. In a linear regression of standardized report on the potential payoff from misreporting, we find that a one dollar increase in incentives changes the standardized report by -0.002 to 0.003 (see Appendix A for more details, in particular how we identify the effect). This means that increasing incentives even further is unlikely to yield the standard economic prediction of +1.

In Appendix A, we show that behavior does not change when subjects report repeatedly. Learning and experience thus do not diminish the effect. Reporting behavior is also quite stable across countries. Country fixed effects are only able to explain 19.7 percent of the between-treatment variation (adjusted $R^2 = 0.088$).

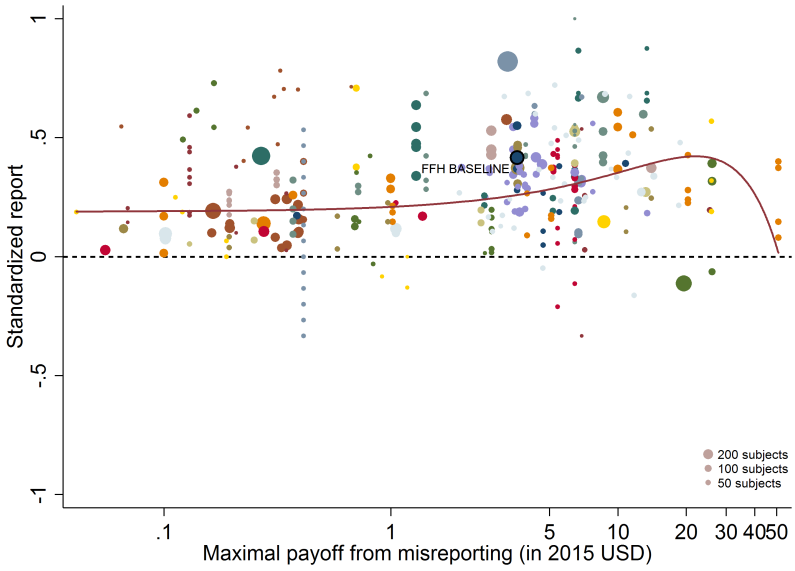
So far, we have focused on the average report. We next analyze the distribution of reports within each treatment.

Finding 3 *All possible reports are made with positive probability.*

Finding 4 *When the distribution of true states is uniform, states that lead to a higher payoff are reported more often.*

Finding 5 *Some non-maximal-payoff reports are made more often than their truthful likelihood to occur.*

Figure 1: Average standardized report by incentive level



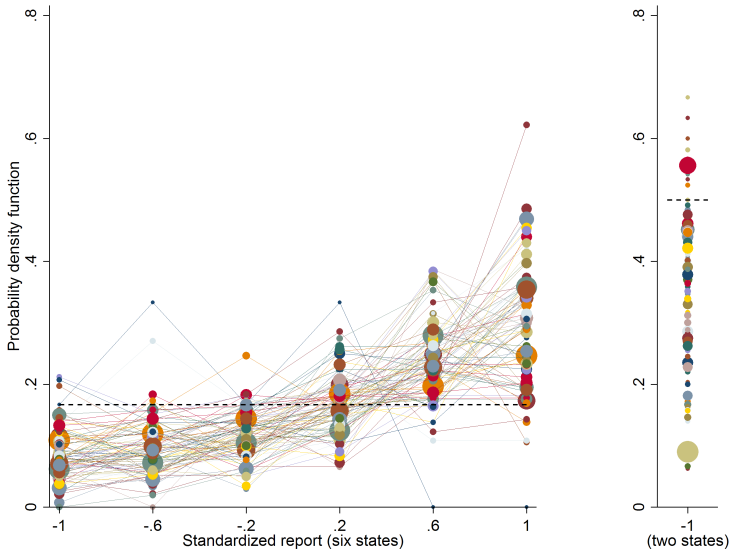
Notes: The figure plots standardized report against maximal payoff from misreporting. Standardized report is on the y-axis. A value of 0 means that subjects realize as much payoff as a group of subjects who all tell the truth. A value of 1 means that subjects all report the state that yields the highest payoff. The maximal payoff from misreporting (converted by PPP to 2015 USD), i.e., the difference between the highest and lowest possible payoff from reporting, is on the x-axis (log scale). Each bubble represents the average standardized report of one treatment and the size of a bubble is proportional to the number of subjects in that treatment. "FFH BASELINE" marks the result of the baseline treatment of Fischbacher and Föllmi-Heusi (2013). The line is the fitted regression line of a quadratic regression.

Figure 2 shows the distribution of reports for all experiments using uniform distributions with six or two states, e.g., six-sided die rolls or coin flips. We exclude the few studies that have non-linear payoff increases from report to report. The figure covers 73 percent of all subjects in the meta study. Each line corresponds to one treatment and the size of the bubbles is proportional to the number of subjects in that treatment. The dashed line indicates the truthful distribution. As one can see, all possible reports are made with positive probability in basically all treatments. The reports that lead to higher payoffs are generally made more often, both for six-state and two-state distributions. The right panel of Figure 2 plots the likelihood of reporting the low-payoff state (standardized report of -1) for two-state experiments. The vast majority of the bubbles are below 0.5 which implies that the high-payoff report is above 0.5.

Interestingly, some reports that do not yield the maximal payoff are reported more often than their truthful probability, in particular the second highest report in six-state experiments is more likely than 1/6 in almost all treatments.

We relegate additional results and all regression analyses to Appendix A.

Figure 2: Distribution of reports (uniform distributions with six and two outcomes))



Notes: The figure depicts the distribution of reports by treatment. The left panel shows treatments that use a uniform distribution with six states and linear payoff increases. The right panel shows treatments that use a uniform distribution with two states. The right panel only depicts the likelihood that the low-payoff state is reported. The likelihood of the high-payoff state is 1 minus the depicted likelihood. The size of a bubble is proportional to the total number of subjects in that treatment. Only treatments with at least 10 observations are included. The dashed line indicates the truthful distribution at 1/6 and 1/2.

2 Theory

We use a unified theoretical framework to formalize various ways to explain the apparent reluctance to lie. One of the main contributions of our paper is to identify five dimensions of behavior along which the models generate distinct behavior. We focus only on those dimensions for our predictions. We find that models differ in (i) how the distribution of true states affects one's report; (ii) how the belief about the reports of other subjects influences one's report; (iii) whether the observability of the true state affects one's report; (iv) whether

some subjects will lie downwards if the true state is observable; (v) whether all reports are made with positive probability. As a prediction (iv'), we also derive whether some subjects will lie downwards if the true state is *not* observable, as in the standard FFH paradigm. We cannot test this last prediction in our data but state it nonetheless as it is helpful in building intuition regarding the models as well as important for potential applications.[8] To cover the breadth of plausible explanations (and to be able to draw robust conclusions), we study a large number of models, in total 22. This includes all models suggested previously in the literature as well as several new models. All models are listed in Table 1 below.

There are three broad types of explanations of why subjects seem to be reluctant to lie: subjects face a lying cost when deviating from the truth; they care about some kind of reputation that is linked to their report; or they care about social comparisons or social norms which affect the reporting decision. In this section, we discuss one example each of these three types of explanations and one of the two models that our data will not be able to falsify. The discussion of all other models is relegated to Appendix B.

We focus on the situation where there are two possible states and two possible reports. Focusing on such a binary setting has several advantages. First, it allows us to abstract from many details of the models. In particular, our predictions are valid independent of the exact specification of individual heterogeneity or the functional form.[9] Second, it allows us to make clear comparative static predictions; for example, in the binary setting, we can derive very precisely the reaction of reports to a change in the distribution of true states. In contrast, with more than two states, the predictions of most models will depend on parameters. Third, for models that rule out downwards lying, the binary setting allows us to back out the full reporting strategy of individuals without actually observing the true state: the high-payoff state will be reported truthfully, so we can deduct the expected number of high-payoff states from the high-payoff reports and we are left with the reports of the low-payoff state. Finally, the majority of experiments surveyed in our meta study indeed uses two states (see Table A.1).[10]

[8] Peer et al. (2014) and Gneezy et al. (2013) study downwards lying in a setting in which at least some subjects will feel unobserved.

[9] For example, for models where individuals care about what others do (e.g., social-comparisons type models) our setting has the feature that it doesn't matter whether individuals care about the average report or the distribution of reports.

[10] One seeming disadvantage is that we cannot capture the over-reporting of non-maximal reports in a binary

To keep notation and intuitions simple, we consider fixed payoffs. In Section 4, we extend the surviving models to accommodate shifts in the payoff size. For all our models, we assume that there is a set of individuals, indexed by k. Each individual observes a state $\omega \in \{\omega_0, \omega_1\}$, which is drawn i.i.d. across individuals, and then takes an action $r \in \{r_0, r_1\}$, which is their report of what their state is. Finally, individuals receive a monetary payment which is equal to their report r. ω is thus the payoff the individual would receive if they reported truthfully. We suppose both the state and action spaces are ordered with $r_1 > r_0$, and that ω_0 corresponds in a natural way to r_0, and ω_1 to r_1. For example, imagine an individual privately flipping a coin. If they report heads, they receive £10, if they report tails they receive nothing. Then $\omega_0 = r_0 = 0$, and $\omega_1 = r_1 = 10$. In the following we will refer to ω_0 and r_0 as the "low" state or report, and to ω_1 and r_1 as the "high" state or report respectively.

We denote a utility function as ϕ (suppressing the index k). For clarity of exposition, we suppose that ϕ is differentiable in all its arguments, except where specifically noted, although our predictions are true even when we drop differentiability and replace our current assumptions with the appropriate analogues. We allow that individuals are heterogeneous: individuals have a type $\theta \in [0, 1] = \Theta$, which will characterize the relative trade off between monetary benefits (which depend only on the report r) and the non-monetary benefits and costs (which will vary depending on the model). Importantly, this means that also the functional form of the utility function can be heterogeneous as long as our assumptions regarding how ϕ depends on r, ω and θ, specified below, are satisfied. We suppose throughout this section that ϕ, and all sub-functionals used in the utility, are continuous in all their arguments (unless otherwise specified). Such continuity is required for the existence of equilibrium when individuals care not just about their own report, but also about the report of others.

We assume that there is a distribution H over θ which is non-atomic. Given a threshold type $\bar{\theta}$, i.e. a type indifferent between the two reports, denote $P(\bar{\theta}) = P(\theta < \bar{\theta})$ the probability of θ being less than threshold type $\bar{\theta}$. We assume that there is a commonly known distribution F over states with full support (with f the corresponding PDF), i.e., $f(\omega_0) \in (0, 1)$. We will describe the induced distribution over reports as G and an individual's belief about G as $\hat{G}$ (with g and $\hat{g}$ as PDF). We assume that individuals only report once and there are no

setting. This behavior can, however, be explained by many different types of models and is thus not very good at distinguishing between models. In Section 4, we show that the model that our data cannot falsify is also able explain non-maximal lying when there are more than two states.

repeated interactions.[11] We suppose for the most part each individual privately observes their own state, except for our third prediction, described below.

In many of our models, a person's utility depends not just on their own action, but also on the actions of others. In those cases, we focus on the Nash Equilibrium of the induced game.[12] For consistency, we also use the word equilibrium to describe the outcome of models in which utilities are not affected by others' actions. A strategy maps type and state combinations ($\theta \times \omega$) into a mixed strategy over reports r, although almost all individuals will play a pure strategy in our framework. This is because all types have measure zero and, given our assumptions on the interaction between θ and the costs of lying in the models we consider (detailed below), if an individual of type $\bar{\theta}$ is indifferent between the two reports, then no other type can be indifferent.

The details of our five testable predictions are as follows. The first three predictions involve comparative statics. The last two involve characteristics of the equilibrium, fixing a set of parameters. This is similar to the stylized facts discussed in the meta study, which are composed both of comparative statics as well as descriptions of characteristics of the equilibrium. Even in our simple binary state/binary report environment some of the models we consider do not guarantee a unique reporting distribution G without additional parametric restrictions. We discuss in more detail how we deal with potential non-uniqueness for each prediction and we mark the models which do not necessarily have unique equilibria with an asterisk (*) in Table 1. Importantly, no model is ruled out solely on the basis of predictions that are based on an assumption of uniqueness. Similarly, the models that cannot be falsified by our data are not consistent solely because of potential multiplicity in equilibria.

Our first prediction tests how the distribution of reports G changes when the high state is more likely to be drawn, i.e., when $f(\omega_1)$ increases. Specifically, we consider changes in the statistic $\frac{f(\omega_0)-g(r_0)}{f(\omega_0)} = 1 - \frac{g(r_0)}{f(\omega_0)}$ starting from the situations where the equilibrium G has full support. Changes to this statistics can be interpreted, with two states, as a drawing in or drawing out effect. With more than two outcomes it is not clear how to define or interpret a similar statistic, another reason why we focus on a setting with only two states. For those models in which no individual lies downwards we can additionally interpret the statistic as

[11] We focus on the case where each individual makes a single report to eliminate standard reputational effects and because it is not clear whether or how individuals aggregate actions in a dynamic, multiple reports setting.

[12] The RH+Cursedness and RH+Level-k models described in Appendix B use other solution concepts.

the proportion of people who draw ω_0 but report r_1, or in other words, the fraction of people who lie upwards.

Definition 1 *We say a model exhibits drawing in if* $1 - \frac{g(r_0)}{f(\omega_0)}$ *is increasing in* $f(\omega_1)$*, drawing out if* $1 - \frac{g(r_0)}{f(\omega_0)}$ *is decreasing in* $f(\omega_1)$*, and f-invariance if* $1 - \frac{g(r_0)}{f(\omega_0)}$ *is not affected by* $f(\omega_1)$.[13]

Our second prediction attempts to understand how an individual's report will change when we exogenously shift their belief $\hat{G}$ about the distribution of reports G. We again focus on situations where G features full support. Such an exercise allows us to test the models in one of two ways. First, for models in which $\hat{G}$ (or a sufficient statistic for it) directly enters the utility, namely all social-comparison models, we can directly assess the effect of a shift in $\hat{G}$ on behavior. For these models, shifting an individual's belief about G directly alters their best response (and since subjects are best responding to their $\hat{G}$, which may be different from the actual G, we may observe out-of-equilibrium behavior).[14]

Second, this exercise allows us, albeit indirectly, to understand what happens when beliefs about H (the distribution of θ, e.g., the idiosyncratic component of lying costs) change. Directly changing this belief is extremely difficult since this requires identifying θ for each subject and then conveying this insight to all subjects. For models with a unique equilibrium, because G is an endogenous equilibrium outcome, shifts in $\hat{G}$ can, however, only be rationalized by subjects as shifts in some underlying exogenous parameter — which has to be H, since our experiment fixes all other parameters (e.g., F and whether states are observable).[15] For many of these models, the conditions defining the unique equilibrium reporting strategy are invariant to shifts in $\hat{G}$ and H, which would mean that our treatment should have not have

[13] In models where the equilibrium is potentially not unique, caution is needed in interpreting the effect of changes in F on behavior. We have two types of predictions. First, for some models the set of possible equilibria is invariant to changes in F. In this case we believe that it is reasonable to assume that our treatment does not induce equilibrium switching and therefore behavior does not change with F. In Table 1 we list these models as exhibiting f-invariance. Second, for other models the set of equilibria changes with changes in F. For these models the predictions of drawing in/out listed in Table 1 are based on the assumption of a unique equilibrium.

[14] These models all feature multiple equilibria, and moreover, they all predict affinity (defined below).

[15] In some models there are additional preference parameters, e.g. the degree of loss aversion. The belief about these parameters will be updated along with H. To specify the updating process more precisely, we suppose that individuals have a single probability distribution H (and a distribution of the other preference parameters) which induces $\hat{G}$ (and G). In a more complete model, individuals would think many different possible H distributions to be possible, and hold a prior over these different distributions. Thus, observing a different $\hat{G}$ would induce a shift in the inferred distribution over the different possible Hs. Given reasonable assumptions about the prior distribution over H and the other parameter distribution our results will continue to hold.

an effect. For another set of models, any given H may give rise to multiple equilibria, so there is no simple mapping from $\hat{G}$ to beliefs about H and we cannot make any prediction about how a shift in $\hat{G}$ should affect G.[16] Last, for one model (the Audit model), there is a simple mapping from H to G and so by changing $\hat{G}$ we can affect the posterior belief about H in a predictable way.

Definition 2 *Given a fixed F, we say a model exhibits affinity if $g(r_1)$ is increasing in $\hat{g}(r_1)$, aversion if $g(r_1)$ is decreasing in $\hat{g}(r_1)$ and g-invariance if $g(r_1)$ is not affected by $\hat{g}(r_1)$.*

Our third prediction is whether or not observability (by outsiders) matters for the distribution of reports. In particular, we will test whether individuals' reports change if the experimenter can observe not only the report, but also the state for each individual.

Definition 3 *We say a model exhibits o-shift if G changes when the true state becomes observable to outside observers, and o-invariance if G is not affected by the observability of the state.*[17]

Our fourth prediction comes in two parts. Both parts try to understand whether or not an individual will engage in downwards lying, i.e., draw ω_1 and report r_0. The first is whether downwards lying will occur in an equilibrium with observability of the state and where G features full support, i.e., the situation described in the third prediction. The second is an analogous prediction but in the situation where the state is not observed by the experimenter. We will only test the former prediction in our experiments.

Our fifth prediction is whether there are parameter values such that an equilibrium exists in which all reports are made with positive probability, i.e., G has full support, when the distribution F is uniform or when the high state has a higher likelihood, i.e., $f(\omega_1) \geq f(\omega_0)$.[18]

[16] These are Reputation for Being Not Greedy, Normalized RH, RH+LC and LC-Reputation.

[17] As for f-invariance, whenever a model has potentially multiple equilibria and this set of equilibria is invariant to observability, we list the model as exhibiting o-invariance because we believe that pure equilibrium switching is unlikely to occur. In contrast to drawing in/out, we do not need to assume a unique equilibrium for o-shift predictions as we do not specify in which direction behavior will move.

[18] Whether a full-support equilibrium exists for $f(\omega_1) < f(\omega_0)$ again depends on the model. For some models, specifically the Standard, Kőszegi-Rabin and Collective Reputation models, there will never be an equilibrium with full support. For other models, specifically the RH, RH+Cursedness, RH+Level-k, and Normalized RH models, there are parameter values so that a full-support equilibrium can exist so long as $f(\omega_1) < f(\omega_0)$.

We summarize the six predictions of all 22 models in Table 1.We also, for comparison purposes, state the results of our experiments in the row labeled Data.[19] All proofs for the models in Section 2 are in Appendix C.

2.1 Standard Model

In the standard model an individual simply maximizes their monetary payoff. Formally, utility depends only on r, not on ω and there is thus no θ, so the utility function is $\phi(r)$ and is strictly increasing in its argument. In the standard model all individuals report r_1 regardless of their state or the observability of their state. Thus, individuals exhibit f-invariance, g-invariance and o-invariance and never lie down. In Appendix B, we discuss the predictions of two popular behavioral models for which, like in the standard model, utility depends on monetary payoffs, and that were not explicitly developed to capture lying aversion, namely inequality aversion (B.1) and the reference-dependent preferences model by Kőszegi and Rabin (B.2).

Several papers (e.g., Demichelis and Weibull 2008, Ellingsen and Östling 2010, Kartik et al. 2014b) assume that individuals have weak (or lexicographic) preferences for truth-telling, i.e., individuals receive an additional small utility $\varepsilon > 0$ when they report truthfully. Since reports in our setup always yield different monetary payoffs, these models predict the same as the standard model.

2.2 Lying Costs (LC)

The first type of explanation for the reluctance to lie assumes that deviating from telling the truth is directly costly to individuals. The fact that individuals' utility also depends on the realized state, not just their monetary payoff, could come from moral or religious reasons; from self-image concerns (if the individual remembers ω and r)[20]; or from injunctive social norms of honesty. Such "lying cost" (LC) models have wide popularity in applications and represent a simple extension of the standard model. Our formulation of this class of models

[19]The Kőszegi-Rabin+LC model has additional predictions which we test and which are not listed in the table. Moreover, our experiments also deliver results for when there are more than two states. We describe additional results for the LC and the RH models in such an environment in Appendix C.1 and C.2.

[20]If the individual forgets about their own state ω and cares about what their own future selves think about them, judging only from their report r (similar to Bénabou and Tirole 2006), then our Reputation for Honesty model, described in the next section, may be more appropriate. Only the predictions regarding observability would need to be adjusted if the audience is "internal". In our setting, given the short length of time between draw of state and report, it seems, however, unlikely that individuals would forget the state but not the report.

Table 1: Summary of Testable Predictions

Model	**Predictions**					**Section**
	Shifts in F	**Shifts in $\hat{G}$**	**Observability**	**Lying Down Unobs./Obs.**	**Full Support**	
Standard Model + Variants						
Standard Model	f-invariance	g-invariance	o-invariance	No/No	No	2.1
Inequality Aversion*	f-invariance	affinity	o-invariance	-/-	Yes	B.1
Kőszegi-Rabin	f-invariance	g-invariance	o-invariance	No/No	No	B.2
Lying Costs (LC)	f-invariance	g-invariance	o-invariance	No/No	Yes	2.2
Reputation						
Reputation for Honesty (RH)	drawing in	g-invariance	o-shift	Yes/No	No	2.3
Reputation for Being Not Greedy*	f-invariance	-	o-invariance	Yes/Yes	Yes	B.4
RH+Cursedness	drawing in	g-invariance	o-shift	Yes/No	No	B.5
RH+Level-k	drawing in	g-invariance	o-shift	Yes/No	No	B.6
Normalized RH*	-	-	o-shift	Yes/No	No	B.7
Collective Reputation	f-invariance	g-invariance	o-invariance	No/No	No	B.8
Social Norms/Comparisons						
Conformity in Lying Costs (CLC)*	drawing out	affinity	o-invariance	No/No	Yes	2.4
Guilt Aversion*	f-invariance	affinity	o-invariance	-/-	Yes	B.9
Inequality Aversion+LC*	drawing in	affinity	o-invariance	-/-	Yes	B.10
Keeping Up with the Joneses*	-	affinity	o-invariance	No/No	Yes	B.11
Conformity in Actions*	f-invariance	affinity	o-invariance	-/-	Yes	B.12
Censored CLC*	-	affinity	o-invariance	No/No	Yes	B.13
Combined Models						
Separable RH+LC*	drawing in	-	o-shift	-/No	Yes	2.5
Kőszegi-Rabin+LC	-	g-invariance	o-invariance	No/No	Yes	B.14
General RH+LC*	-	-	o-shift	- /No	Yes	B.15
LC-Reputation*	-	-	o-shift	- /-	Yes	B.16
Audit	drawing in	aversion	o-shift	No/No	Yes	B.17
CLC+LC*	drawing out	affinity	o-invariance	No/No	Yes	B.18
Data	**drawing in**	**g-invariance**	**o-shift**	**?/No**	**Yes**	

Notes: The details of the predictions are explained in the text. "-" means that no specific prediction can be made. Models which do not necessarily have unique equilibria are marked with an asterisk (*). For these models, the predictions of f-invariance and o-invariance mean that the set of possible equilibria is invariant to changes in F or observability. The predictions of drawing in/out are based on the assumption of a unique equilibrium.

nests all of the lying cost models discussed in the literature, including a fixed cost of lying, a lying cost that is a convex function of the difference between the state and the report, and generalizations that include different lying cost functions.[21]

Formally, we suppose individuals have a utility function

$$\phi(r, c(r, \omega); \theta)$$

c is a function that maps to the (weak) positive reals and denotes the cost of lying. We suppose that c has a minimum when $r = \omega$, so that individuals experience no cost when they tell the truth, i.e., we normalize $c(\omega, \omega) = 0$. We make a few assumptions on ϕ. First, ϕ is strictly increasing in the first argument (r), fixing all the other arguments; this captures the property that utility is increasing in the monetary payment received. Second, ϕ is weakly decreasing in the second argument, fixing all the other arguments, capturing the property that utility falls as the cost of lying increases. In particular, if $\theta = 0$, the partial of ϕ with respect to the second argument is 0 (so that we nest the standard model). Third and fourth, fixing all other arguments, ϕ is decreasing in θ, and the cross partial of ϕ with respect to c and θ is negative (strictly so when both arguments are strictly positive). This captures the properties that an individual with a higher draw of θ has both a higher utility cost of lying, for the same "sized" lie, and faces a higher marginal cost of lying (in other words utility exhibits increasing differences with respect to c and θ).[22] Our predictions regarding the LC model do not depend on individuals all having the same functional form c so long as the assumptions regarding θ hold. So, for example, our results hold when some individuals have fixed and others convex costs of lying.

[21] This includes, for example, Ellingsen and Johannesson (2004); Kartik (2009); Fischbacher and Föllmi-Heusi (2013); Gibson et al. (2013); Gneezy et al. (2013); Conrads et al. (2013); Conrads et al. (2014); DellaVigna et al. (2014); and Boegli et al. (2016).

[22] For some specifications, for example fixed costs of lying, c will not be differentiable in its arguments (but our results still will hold if we replace our current assumptions with their appropriate analogues). Here, and in the rest of the models, we do not need that utility is decreasing in θ, only that the restriction on the cross partials hold. We make this restriction as it allows for a natural interpretation of θ. Given that there are only two possible states and reports, and the property that people do not lie downwards in the LC model, we only need the restriction that higher θs generate a higher utility cost of lying. We have the more general restriction in order to make the comparison of the assumptions across models easier. Finally, we could consider the functional form $\phi(r, \omega; \theta)$. Such a functional form nests the model described in the body of the text. We use the more specific functional form because it already nests the lying cost models used in the literature and allows for a simpler exposition of the natural assumptions on utility that the literature ascribes to lying cost models.

The LC model predicts the following. Because individuals (weakly) pay a cost of lying downwards, and also receive a lower monetary payoff when reporting r_0, they will never lie downwards. Since only their own state and their own report matter for utility, conditional on drawing the low state, for a fixed θ, an individual will always make the same report, regardless of F or $\hat{G}$. Thus, we observe both f-invariance and g-invariance. The f-invariance result (suitably interpreted) generalizes to a setting with many states, see Appendix C.1 for details. Moreover, if the lying costs are large enough, some individuals will not want to lie up, generating full support for any distribution F. Last, the lying cost is an internal cost and does not depend on the inference others are making about any given person. Thus, individuals do not care whether their state was observed.

In Section 2.5 and in Appendix B (in particular B.10, B.14, B.15, B.16, B.17 and B.18), we discuss several models in which individuals care about lying costs and also about reputation or social norms (including the special case of a mixture model in which some individuals care about lying costs and others care about, say, reputation).

2.3 Reputation for Honesty (RH)

The second type of explanation posits that individuals care about some kind of reputation that is linked to their report. In this section, we assume that individuals care about how honest they appear to others (regardless of their actual honesty). In the Reputation for Honesty (RH) model, an individual's utility is falling in how likely it is that a randomly drawn individual making that report is a liar, i.e., has a state not equal to the report. Akerlof (1983) provides the first discussion in the economics literature of this motivation for honesty and many recent papers have built on this intuition.[23] One reason why individuals might care about the appearance of honesty is due to a heuristic that many interactions, unlike our experiment, are repeated and non-anonymous, and so developing a reputation for honesty is valuable per se.

Formally, individuals' utility is

$$\phi(r, \Lambda(r); \theta)$$

[23]This includes, for example, Mazar et al. (2008); Suri et al. (2011); Hao and Houser (2013); Shalvi and Leiser (2013); Utikal and Fischbacher (2013); Fischbacher and Föllmi-Heusi (2013); Gill et al. (2013) and Hilbig and Hessler (2013).

$\Lambda(r)$ is the fraction of liars, i.e. individuals who report r but drew a state $\omega \neq r$. We assume (as in LC) ϕ is strictly increasing in its first argument and decreasing in the second argument; if $\theta = 0$, the partial of ϕ with respect to the second argument is 0. These assumptions capture the property that individuals prefer a higher monetary payoff but dislike being thought of as a liar. Moreover, we suppose that ϕ is decreasing in θ fixing the first two arguments, and that the cross partial of ϕ with respect to $\Lambda(r)$ and θ is negative (strictly so when both arguments are strictly positive). This again captures the property that a higher draw of θ indicates both a higher utility cost for any given reputation and a higher marginal cost.

RH predicts that, first, because utility depends only on reports but not on the drawn state, individuals' best response function will depend only on their individual θ, not on the drawn state. This means that conditional on having the same θ, individuals drawing the low or the high state will make the same report. Thus, in any equilibrium that features full support, some individuals must lie downwards. Observe that starting with an equilibrium with full support, there is an indifferent type: an individual who is indifferent between the high and low reports. Everyone with a lower θ gives the high report; everyone with a higher θ gives the low report. We call the former probability $P(\bar{\theta})$. Thus the fraction of liars at the high report is $\Lambda(r_1) = \frac{P(\bar{\theta})f(\omega_0)}{P(\bar{\theta})f(\omega_0)+P(\bar{\theta})[1-f(\omega_0)]} = f(\omega_0)$. Similarly, we can show that $\Lambda(r_0) = f(\omega_1)$. This implies directly that if $f(\omega_0) \leq f(\omega_1)$ then in an equilibrium with full support the fraction of liars at r_1 would be smaller than the fraction of liars at r_0, and so by saying r_1, individuals would receive both a higher monetary payoff and a lower reputational cost. Thus, all individuals should say r_1 and there cannot be an equilibrium with full support. This result generalizes to a setting with many states, see Appendix C.2 for details.

Now, consider what happens when $f(\omega_1)$ increases, starting from a full-support equilibrium. Intuitively, because there are more individuals who initially drew ω_1 the fraction of liars at r_1 decreases. Thus, reporting r_1 becomes more attractive, and so individuals who were initially indifferent between reporting the high and low report now strictly prefer to report the high report. Thus we observe drawing in.[24]

Next, fixing F, consider what happens when the individual's $\hat{G}$ changes. Surprisingly, and

[24]Because Λ (as well as G) is endogenous, with more than two states/reports there may be complicated shifts in Λ and G due to even simple changes in F. However, with binary states the comparison is clear-cut; and as we discussed previously we conceive of drawing in and out as a test with two states/reports.

importantly, the Λs derived in the previous paragraph do not depend on G or H.[25] Therefore, both the costs and benefits of any given report do not change with $\hat{G}$.

Because in the RH model others make inferences about whether an individual lies, it is important whether the state is observable or not. If only the report is observable then they must infer the probability of being a liar from the report. If also the state is observable, lying can be directly identified and the probability of being a liar is either 0 or 1. If the state is observable, individuals with low θs will give the highest report, and those with high θs will state the truth (regardless of the number of states).

We also consider a range of other reputation models. Section 2.5 describes a model that combines lying costs and a reputation for honesty. In Appendix B, we discuss a model in which the reputation is about being not greedy (B.4); two models that relax the rationality of the audience, building on cursedness (B.5) and level-k reasoning (B.6); a model in which the reputation cost is normalized by the average reputation cost in equilibrium (B.7); a model in which individuals care about the collective reputation of all subjects (B.8); a model in which individuals have lying costs and care about the reputation about their lying cost parameter θ (B.16); and a model in which the reputation is about having lied upwards (B.17). All results are summarized in Table 1.

2.4 Social Norms: Conformity in Lying Costs (CLC)

The third type of explanation posits that individuals care about social norms or social comparisons which inform their reporting decision. The leading example is that individuals may feel less bad about lying if they believe that others are lying, too. In other words, if there is a social norm of lying, lying incurs a lower utility cost. Importantly, these norms are "descriptive" in the sense that they depend on the behavior of others, rather than injunctive, where they do not depend on what others are doing (injunctive norms are better captured by LC models). We call such a model conformity in lying costs (CLC). Such concerns for social norms are discussed, for example, in Gibson et al. (2013), Rauhut (2013) and Diekmann et al. (2015). Our model follows the intuition of Weibull and Villa (2005).

[25] This is not to say that G and Λ are not related; another way of writing $\Lambda(r_1)$ is $\frac{P(\bar{\theta})f(\omega_0)}{G(r_1)}$. The independence of Λ from G and H is true only in equilibrium, where all individuals are best responding to their belief $\hat{G}$. If some individuals fail to take their best-response action, this is no longer true.

Formally, in the CLC model individuals have a utility function

$$\phi(r, \eta(c(r, \omega), \bar{c}); \theta)$$

$c(r, \omega)$ has the same interpretation and assumptions as in the LC model. $\bar{c}$ is the average incurred cost of lying in society. This average lying cost is determined in equilibrium, and thus all individuals know what it is. η captures the "normalized cost of lying", i.e. the cost of lying conditional on the incurred lying costs of everyone else. We suppose $\eta(0, \bar{c}) = 0$. For $c > 0$, η is strictly increasing in the first argument and strictly falling in the second so that the normalized cost is increasing in the individual's own personal lying cost and falling in the aggregate lying cost, i.e., their lying costs are falling as others lie more. As in the previous models ϕ is strictly increasing in its first argument, fixing the second argument, and weakly decreasing in the second argument, fixing the first argument. If $\theta = 0$, the partial of ϕ with respect to the second argument is 0. ϕ is decreasing in θ fixing the first two arguments, and the cross partial of ϕ with respect to η and θ is negative (strictly so when both arguments are strictly positive). These assumptions are analogous to the ones presented in the previous models and capture the same intuitions.

The CLC model predicts the following. First, as in the LC model, individuals will never lie downwards since they would face a lower monetary payoff as well as a weakly higher cost of lying. Second, consider what happens as $f(\omega_1)$ increases. Suppose the equilibrium reporting distribution is unique. As $f(\omega_1)$ increases, more individuals draw the high state and can report r_1 without having to lie. Thus, the average cost of lying falls. This increases the normalized cost of lying (η) for all individuals. Thus, an individual who draws ω_0, and was indifferent before between r_0 and r_1 will now strictly prefer r_0. Thus, we observe drawing out. If the equilibrium is not unique, we cannot make a specific prediction.

Although the observed reporting distribution may not be unique, because G enters directly into the utility function we can tell how the individual's best response changes with shifts in expected G, i.e. $\hat{G}$. Fixing F, if $\hat{g}(r_1)$ increases, more people draw the low state but say the high report. This means that more individuals are expected to lie, and so the normalized cost of lying (η) decreases. Thus, individuals who draw the low report will be more likely to say the high report. Thus we will observe affinity. Last, as in the LC model, these costs do

not depend on any inference others are making, and so individuals do not care whether their state was observed.

In Appendix B, we discuss several other ways of formalizing the effect of social norms and social comparisons on reporting. We discuss a model of inequality aversion (B.1); a model that combines lying costs with inequality aversion (B.10); a model in which individuals care about not falling behind others and where the true state is their reference point (B.11); a model in which individuals do not care about conformity in lying costs but about conformity in reports (B.12); a CLC model in which only subjects who could have lied upwards matter for conformity (B.13); and a combination of lying costs and CLC (B.18). We also discuss the well-known model of guilt aversion under this header (B.9) as it turns out to be very similar to an inequality aversion model in our setting. All results are summarized in Table 1.

2.5 Separable Reputation for Honesty plus Lying Costs (RH+LC)

It is plausible (albeit less parsimonious) that multiple motives drive the reluctance to lie (e.g., Khalmetski and Sliwka 2016). In this section, we consider a model in which individuals may experience both an intrinsic cost of lying, as well as reputational costs associated with inference about their honesty, which we call the Separable RH+LC model. Since lying costs are our preferred way to capture self-image concerns about honesty, one possible interpretation of this model is that individuals care about self-image and social image. We focus on a situation where there is additive separability between the different components of the utility function. In Appendix B.15, we consider the General RH+LC model which allows for non-separability but makes weaker predictions. Formally, in the Separable RH+LC model let utility be

$$\phi(r, c(r,\omega), \Lambda(r); \theta_1, \theta_2) = u(r) - \theta_1 c(r,\omega) - \theta_2 \upsilon(\Lambda(r))$$

u is strictly increasing in r. c is as described in the LC-model: weakly positive with a minimum at $r = \omega$ and $c(\omega, \omega) = 0$. υ is a strictly increasing function of $\Lambda(r)$, the fraction of liars at r. υ is weakly positive and $\upsilon(0) = 0$. Thus the individual likes more money, but dislikes lying and being perceived as a liar. Because there are two dimensions of costs, there is now a joint distribution H over $\theta_1 \times \theta_2$ which is non-atomic.[26]

[26] If we suppose that H may be atomic, then we can also capture "mixture" models, where each individual

The predictions are as follows. Because individuals have a concern for reputation and also have lying costs, they may or may not lie down. If the former is a strong motivation, some individuals will lie down. If the lying costs dominate, they will not. Unlike the RH model, individuals will have different best response functions depending on their draw, because of the lying costs. Importantly, and distinct from either the LC or the RH models individually, multiple equilibria may occur.

The Separable RH+LC model exhibits drawing in whenever the equilibrium is unique. In order to gain intuition for this, consider what happens when $f(\omega_1)$ increases. First, some individuals who previously drew ω_0 will now draw ω_1. Those individuals now face a lower lying cost when giving the high report. Suppose some of those individuals actually now give the high report. Fixing the behavior of others, this also reduces the fraction of liars giving the high report and the reputational cost of r_1 decreases; and similarly, increases the fraction of liars giving the low report. Thus, it reduces the (relative) cost of giving the high report even more. Therefore, we observe drawing in. Of course, such intuition relies on partial equilibrium reasoning, but the formal proof in Appendix C.4 shows how to extend this to full equilibrium reasoning.

Since equilibria are not necessarily unique, we cannot predict the effect of our $\hat{G}$ treatments.[27] Because the model includes reputation costs, whether or not outsiders observe just the report, or also the state, matters for behavior, just as in the RH model alone. Moreover, because the motivation to lie downwards is driven by reputational concerns, when the states are observed, individuals do not lie down.

---

either only cares about lying costs, or only cares about reputational costs, but there is a mix in the total population. In this case, H would have zero support everywhere where both θs are strictly greater than 0.

[27] Even with a unique equilibrium, we may observe either aversion, affinity or g-invariance since it depends on how the distribution of H is perceived to have changed by individuals when $\hat{G}$ shifts. If, for example, the change is interpreted as a shift by individuals who have low reputational costs, and so care mostly about lying costs, then an increase in $\hat{g}(r_1)$ will be interpreted as more individuals who drew ω_0 being willing to give the high report. This decreases the proportion of truth-tellers at the high report, driving aversion. In contrast, suppose the change is interpreted as a shift by individuals who have medium lying costs, but relatively high reputational costs. This means that it is interpreted as a shift in the reports of individuals who drew the high state (since individuals who drew the low state and have medium lying costs are unlikely to ever give the high report). An increase in $\hat{g}(r_1)$ is then interpreted as individuals who drew ω_1 as being more willing to pay the reputation cost of reporting r_1. Thus, the fraction of truth-tellers at r_1 increases, driving affinity.

3 New experiments

In this section we report a large-scale (N = 1610) set of experiments designed to test the predictions outlined above. The experiments were conducted with students at the University of Nottingham and University of Oxford.[28] The experiments focus on three empirical tests regarding: i) the effects of shifting the distribution of true states F; ii) the effects of shifting beliefs about the distribution of reports $\hat{G}$; and iii) the effects of observability of the true state ω. Where appropriate, we will also discuss how the data collected in the new experiments can shed light on the predictions regarding full support of reports and downward lying.

3.1 Shifting the distribution of true states F

We test the effect of a shift in the distribution of true states F in two sets of treatments. The first set of treatments uses binary distributions of true states, as in the theoretical analysis developed in the previous section. Subjects are invited to the laboratory for a short session in which they are asked to complete a questionnaire that contains some basic socio-demographic questions as well as filler questions about their financial status and money-management ability. Subjects are told that they would receive money for completing the questionnaire and that the exact amount would be determined by randomly drawing a chip from an envelope. The chips have either the number 4 or 10 written on them, representing the amount of money in GBP that subjects are paid if they draw a chip with that number. Thus, drawing a chip with 4 on it represents drawing ω_0 and drawing a chip with 10 represents drawing ω_1. Reports of 4 and 10 are similarly r_0 and r_1. The chips are arranged on a tray on the subject's desk such that subjects are fully aware of the distribution F (see Appendix D for a picture of the lab setup). Subjects are told that at the end of the questionnaire they need to place all chips into a provided envelope, shake the envelope a few times, and then randomly draw a chip from the envelope. They are told to place the drawn chip back into the envelope and to write down the number of their chip on a payment sheet. Subjects are then paid according to the number reported on their payment sheet by the experimenter who has been waiting outside the lab for the whole time.

[28] Subjects were recruited using ORSEE (Greiner 2015). The computerized parts of the experiments were programmed in z-Tree (Fischbacher 2007). All instructions and questionnaires are available in Appendix D.

We conduct two between-subject treatments, varying the distribution of chips that subjects have on their trays. In one treatment the tray contains 45 chips with the number 4 and 5 chips with the number 10. In the other treatment the envelope contains 20 chips with the number 4 and 30 chips with the number 10. We label the two treatments F_LOW and F_HIGH respectively to indicate the different probabilities of drawing the high state (10 percent vs. 60 percent). We select samples sizes so that the expected number of low states is the same (and equal to 131) in the two treatments. Thus, we have 146 subjects in F_LOW and 328 subjects in F_HIGH. Most of the sessions were conducted in Nottingham and some in Oxford between June and December 2015. Reports are not different across locations (p = 0.359; OLS with robust SE), and we thus report aggregate results pooling data from the two labs.

The second set of treatments uses 10-state distributions. The setup is identical to before except that the tray contains chips numbered 1 to 10. In one treatment (F10_LOW) the tray contains 5 chips with each of the numbers 1–6, 17 chips with the number 7, and 1 chip with each of the numbers 8, 9 and 10. In the other treatment (F10_HIGH) the tray contains 5 chips with each of the numbers 1–6, 1 chip with each of the numbers 7, 8 and 9, and 17 chips with the number 10. Note that the left tails of the distributions (i.e. the probabilities of numbers 1–6) are identical across the two treatments. The two treatments differ in the right tail of the distribution and in particular in the probability mass at 7 and 10. Most models make ambiguous predictions for the F10 treatments. The LC model, however, predicts that there will be no difference in the fraction of subjects reporting numbers 7–10 (see Observation 18 in Appendix C). These experiments were conducted in Nottingham between May and June 2015 with a total of 284 subjects.

3.2 Results

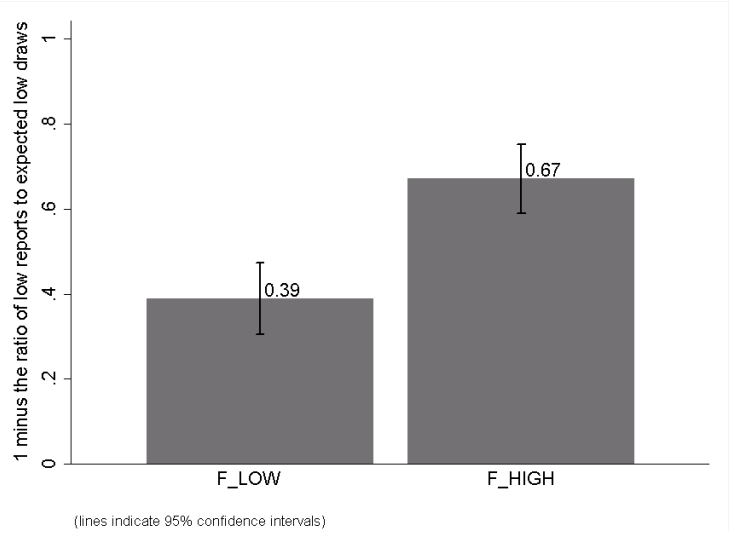
Our first result concerns the binary distribution experiments.

Finding 6 *We observe drawing in, i.e., the statistic* $1 - \frac{g(r_0)}{f(\omega_0)}$ *is significantly higher in F_HIGH than F_LOW.*

Figure 3 shows the values of the statistic $1 - \frac{g(r_0)}{f(\omega_0)}$ across the two treatments. In F_LOW we expect 131 subjects to draw the low £4 payment and we observe 80 subjects actually reporting 4, i.e. our statistic is equal to $1 - \frac{80}{131} = 0.39$. In F_HIGH we also expect 131

subjects to draw 4, but only 43 subjects report to have done so, so our statistic is equal to 0.67.[29] This difference of almost 30 percentage points is very large and highly significant ($p < 0.001$; OLS with robust SE).

Figure 3: Effect of shifting the distribution of true states



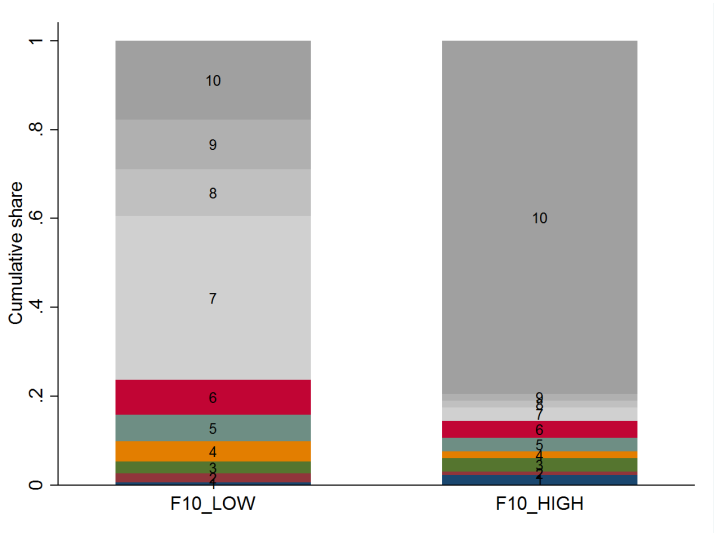
The F10 treatments also yield a treatment difference:

Finding 7 *The proportion of subjects reporting a number between 7 and 10 is significantly higher in F10_HIGH than F10_LOW.*

Figure 4 shows the distribution of reports across the two treatments with 10-state distributions. Recall that the left tail of the distributions of true states is identical across the two treatments. Figure 4 shows, however, that more subjects report 7 to 10 in F10_HIGH than F10_LOW (86 percent vs. 76 percent, p= 0.045, OLS with robust SE). Thus, shifting the probability of high outcomes in the right tail of the distribution draws in subjects from the left tail of the distribution.

[29] This means that 45 percent of subjects in F_LOW and 87 percent in F_HIGH report 10.

Figure 4: Distribution of reports in F10_LOW and F10_HIGH



Treatment F_HIGH provides a further test. As described in Section 2, some models, e.g., RH, predict no full support of reports if $f(\omega_1) \geq f(\omega_0)$. The meta study shows conclusively that this is not in line with the data but this uses only uniform distributions, i.e., $f(\omega_1) = f(\omega_0)$. Treatment F_HIGH shows that there is also full support of reports for a distribution of true states with the probability of the high state strictly higher than the probability of the low state ($f(\omega_1) > f(\omega_0)$). In F_HIGH 13 percent of subjects report the low £4 payment.[30]

Gneezy, Kajackaite, and Sobel (personal communication) and Garbarino, Slonim and Villeval (personal communication) have run FFH-type experiments in which they vary the prior probability of the most profitable state. Similar to our findings in the F10 treatments, Gneezy et al. observe an increase in the frequency of non-maximal reports when the probability of the most profitable state decreases. Garbarino et al. find a similar drawing-in effect as we do.

[30] Shalvi et al. (2011) and Gächter and Schulz (2016b) suggest that subjects might misinterpret the instructions, perhaps deliberately so, and draw twice or three times to then report the highest of the draws. Indeed, in our instructions subjects were told to draw "a" chip from the envelope rather than "one and only one" chip. This heuristic matches the data in F_HIGH quite well (share of 10s: 0.87 (data) vs. 0.84 (best-of-two) or 0.94 (best-of-three)). However, it neither matches the data of F_LOW (0.45 (data) vs. 0.19 (best-of-two) or 0.27 (best-of-three)) nor does it match the data from the meta study, in particular for binary distributions (share of high report: 0.58 (data) vs. 0.75 (best-of-two) or 0.87 (best-of-three)).

3.3 Shifting beliefs about the distribution of reports $\hat{G}$

Our next set of treatments is designed to test predictions concerning the effects of a shift in subjects' beliefs about the distribution of reports, i.e., $\hat{G}$. Our strategy to shift beliefs is based on an anchoring procedure (Tversky and Kahneman 1974): we ask subjects to think about the behavior of hypothetical participants in the F_LOW experiment and we anchor them to think about participants who reported the high state more or less often. Our design has the advantage that we do not show to subjects the actual past behavior of other subjects. Showing actual behavior might trigger norms of conformity and not just affect beliefs. Moreover, if the past behavior is highly selected but presented as if representative, it could be judged as implicitly deceiving subjects and could confound results of an experimental study on deception. Finally, our design minimizes the risk of experimenter demand. We are not aware of other studies that have used anchoring to affect beliefs before.

In our setup, subjects are asked to read a brief description of a "potential" experiment which follows the instructions used in the F_LOW experiment, i.e., 90 percent probability of the low payment and 10 percent probability of the high payment. Subjects also have on their desk the tray with chips and envelope that subjects in the F_LOW experiment had used. Subjects are then asked to "imagine" two "possible outcomes" of the potential experiment. There are two between-subject treatments, varying the outcomes subjects are asked to imagine. In treatment G_LOW the outcomes have 20 percent and 30 percent of hypothetical participants reporting to have drawn a 10, while in treatment G_HIGH these shares are 70 percent and 80 percent. Subjects are then asked a few questions about these outcomes.[31] Subjects are then told that the experiment has actually been run in the same laboratory in the previous year and they are asked to estimate the fraction of participants in the actual experiment who have reported a 10. Subjects are paid £3 for making a correct estimate (within an error margin of at most 3 percentage points). We use these estimates to check whether our anchoring manipulation is successful in shifting subjects' beliefs.

Finally, after answering a few additional socio-demographic questions, subjects are told

[31] Subjects are first asked to compute the truthful chance of drawing a 10 in the potential experiment. For each of the imagined outcomes, they are then asked to estimate how many of the hypothetical participants who report a 10 have truly drawn a 10 as well as questions about what could motivate someone who has drawn a 4 to report either truthfully or untruthfully. Subjects are then asked to rate the satisfaction of someone who reports either a 4 or a 10 in the potential experiment. Finally subjects are asked to estimate which of the two imaginary outcomes shown to them they think is "more realistic".

that they will be paid an additional amount of money on top of their earnings from the belief elicitation. To determine how much money they are paid, subjects are asked to take part in the F_LOW experiment themselves. The procedure is identical to the description of F_LOW in the previous section. The experiments were conducted in Nottingham between March and May 2016 with a total of 340 subjects (173 in G_LOW, 167 in G_HIGH).

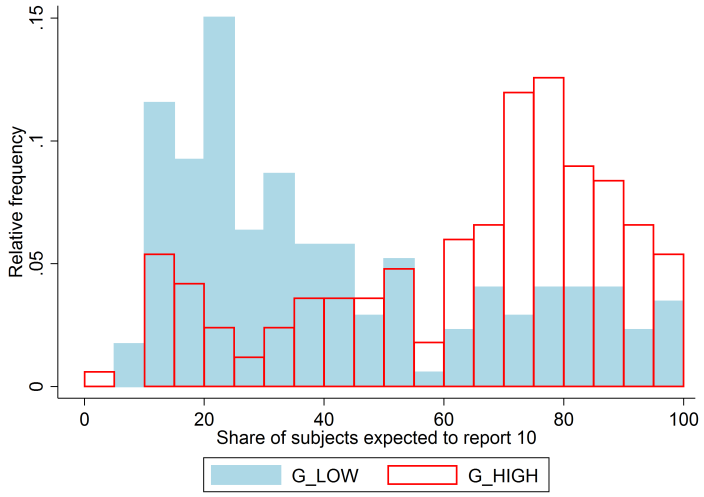
3.4 Results

We start by showing the effect of the anchors on subjects' beliefs.

Finding 8 *The anchors significantly shift beliefs. Estimates of the fraction of participants reporting a 10 are more than 20 percentage points higher in G_HIGH than G_LOW.*

Figure 5 shows the distributions of estimates of the proportion of reported 10s made by subjects across the two treatments. The distribution of the G_HIGH treatment is strongly shifted to the right relative to G_LOW. On average, subjects in G_LOW believe that 41 percent of participants in the F_LOW experiment have reported a 10. In G_HIGH the average belief is 62 percent ($p < 0.001$, OLS with robust SE).

Figure 5: Distribution of beliefs about proportion of reported 10s



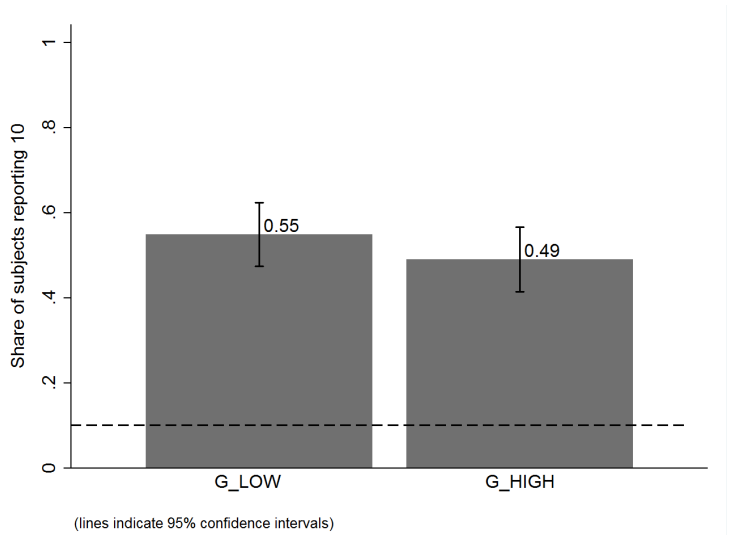
Having established that our manipulation is successful in shifting beliefs about reports in the expected direction, our next step is to examine the effects of this shift in beliefs on

subjects' actual reporting behavior.

Finding 9 *The fraction of subjects reporting a 10 is not significantly different between G_HIGH and G_LOW, i.e., we find g-invariance.*

Figure 6 shows the share of subjects reporting a 10 across the two treatments. Recall that in both treatments the true probability of drawing a 10 is 10 percent (this is indicated by the dashed line in the figure). We observe 55 percent of subjects reporting a 10 in G_LOW, and 49 percent in G_HIGH. This difference is not significant (p = 0.285, OLS with robust SE; p= 0.311, 2SLS regressing report on belief with treatment as instrument for belief).[32]

Figure 6: Effect of shifting beliefs about the distribution of reports



3.5 Changing the observability of states

A final set of treatments tests whether the experimenter observing the subject's true state affects reporting behavior. The experiments use a setup similar to the one used for the

[32]There are only few other studies testing the causal effect of beliefs on reporting, all of them presenting past behavior of other subjects. Diekmann et al. (2015) and Gächter and Schulz (2016a) find no effect and Rauhut (2013) finds a positive effect. However, Rauhut (2013) compares subjects who have initially too high beliefs and are then updated downwards to subjects who have initially too low beliefs and are updated upwards. The treatment is thus not assigned fully randomly. A few studies also find a positive correlation between beliefs about other subjects' reports and the own report (e.g., Abeler et al. 2014). This evidence should, however, be treated with caution, since we also find a positive and significant correlation in our study even though the causal effect of beliefs is zero: In G_LOW the subjects who report a 10 believe that 49 percent of previous participants also reported a 10, whereas the subjects who report a 4 believe that only 30 percent reported a 10 in the previous experiment. The respective numbers in G_HIGH are 72 and 53 percent. In both cases the differences in beliefs are highly significant (Mann-Whitney tests; both $p < 0.001$).

F10 treatments described above. Subjects are invited to the lab to fill in a questionnaire and are paid based on a random draw that they perform privately. There are two between-subject treatments. In both treatments the draw is performed out of a 10-state uniform distribution. In our UNOBSERVABLE treatment, the draw is performed using the same procedures described for the previous experiments: subjects draw a chip at random out of an envelope, report the outcome on a payment sheet, and are paid based on this report. Thus, in this treatment the experimenter cannot observe the true state of a subject and cannot tell for any individual subject whether they lie or tell the truth.

In our OBSERVABLE treatment we maintain this key feature of the FFH paradigm, but make subjects' true state observable to the experimenter. In order to do so, the procedure of the OBSERVABLE treatment differs from the UNOBSERVABLE treatment in two ways. First, the draw is performed using the computer instead of the physical medium of our other experiments (the chips and the envelope).[33] Second, we introduce a payment procedure that makes it impossible for the experimenter to link a report to an individual subject. Before the start of the experiment the experimenter places an envelope containing 10 coins of £1 each on each subject's desk. Subjects are told to sit "wherever they want" and sit down unsupervised. The experimenter does thus not known which subject is at which desk. After the computerized draw, instead of writing the number on their chip on the payment sheet, subjects are told to take as many coins from the envelope as the number of their chip. Subjects then leave the lab without signing any receipt for the money taken or meeting the experimenter again. At the end of the experiment, the experimenter counts the number of coins left by subjects on each desk to reconstruct their "report" and compares it to the true state drawn on the corresponding computer without being able to link any report to the identity of a subject.[34] We ran these

[33]The computerized program simulates the process of drawing a chip from an envelope. Subjects first see on their screen a computerized envelope containing 50 chips numbered between 1 and 10. Subjects have to click a button to start the draw. The chips are then shuffled in the envelope for a few seconds and then one chip at random falls out of the envelope. Subjects are told that the number of that chip corresponds to their payment amount. For comparability, the computer is also used in the UNOBSERVABLE treatment where subjects use it to get precise information on how to perform the (physical) draw.

[34]Had we only introduced observability of states without the double-blind payment procedure, we would have deviated from the FFH paradigm whereby an individual cannot be caught lying. This could confound the results because additional concerns may have come to the fore in subjects' mind. For instance, they may have become concerned with material punishment for misreporting their draw (e.g. exclusion from future experiments). As a robustness check, we invited an additional 69 subjects to participate in a version of the OBSERVABLE treatment that did not use the double-blind payment procedure. The share of subjects misreporting their draw is lower when we do not use the double-blind payment procedure though this effect is not significant.

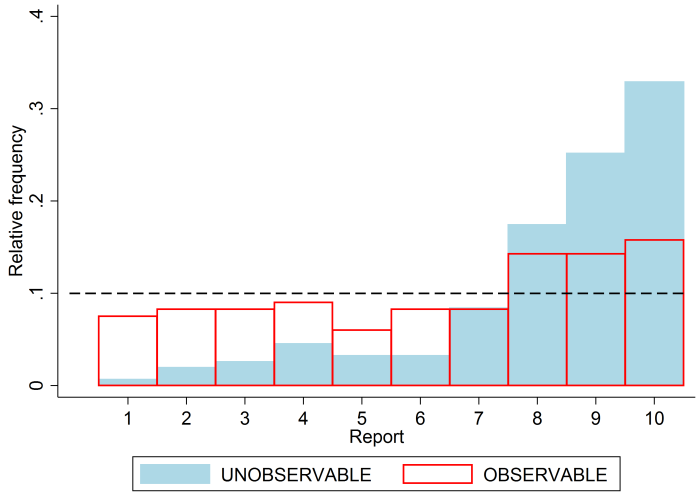
experiments at the University of Nottingham with 288 subjects (155 in UNOBSERVABLE; 133 in OBSERVABLE). Experiments were conducted between May and October 2015.

3.6 Results

Figure 7 shows the distribution of reports in the UNOBSERVABLE and OBSERVABLE treatments. The dashed line in the figure indicates that in both treatments the truthful probability of drawing each state is 10 percent.

Finding 10 *Introducing observability has a strong and significant effect on the distribution of reports.*

Figure 7: Effect of changing the observability of states



Reports in the UNOBSERVABLE treatment are considerably higher than in the OBSERVABLE treatment ($p < 0.001$ Kolmogorov-Smirnov test, $p < 0.001$ OLS with robust SE; see Kajackaite and Gneezy (2015) for a similar result).

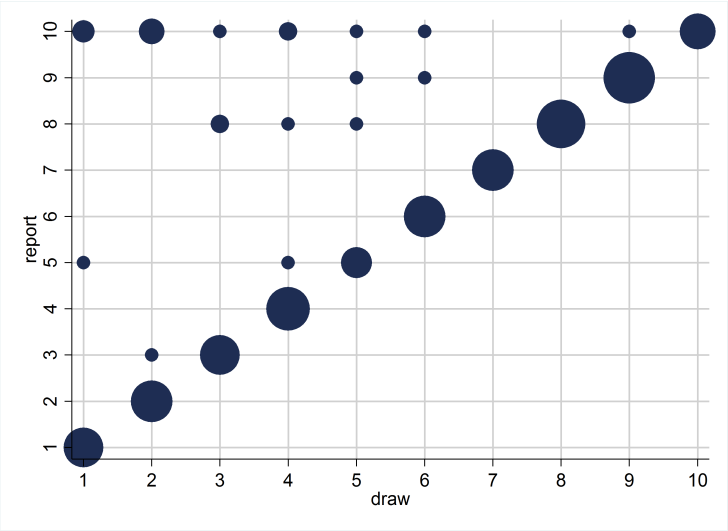
This result also demonstrates that it would be misleading to rely on evidence from settings in which the true state is observable by the researcher if one is actually interested in understanding a setting in which the true state is truly unobservable.

We can also use the OBSERVABLE treatment to examine our prediction about the existence of downward lying.

Finding 11 *We do not find any evidence of downward lying when the true state is observable.*

Figure 8 shows a scatter plot of subjects' reports and true draws in the OBSERVABLE treatment. The size of the bubbles reflects the underlying number of observations. No subject reported a number lower than their true draw, i.e. lied downwards. About 60 percent of the subjects who lie report the highest possible number, the remaining 40 percent of liars report non-maximal numbers.

Figure 8: Reports and true draws in OBSERVABLE



4 Relating theory to data

In this section, we compare the predictions derived in Section 2 and Appendix B with our experimental results and show that only two closely-related models are able to explain the data. We then discuss a simple, parameterized utility function for one of the surviving models which is able to quantitatively reproduce the data from the meta study as well as those from our experiments.

Recall that our five empirical tests concern (i) how the distribution of true states affects one's report (we find drawing in); (ii) how the belief about the reports of other subjects influences one's report (we find g-invariance); (iii) whether the observability of the true state affects one's report (we find it does); (iv) whether some subjects will lie downwards if the true state is observable (we find they don't); (v) whether all reports are made with positive

probability (we find they are).[35] Table 1 summarizes the predictions of all models. One can see that only the RH+LC models and the LC-Reputation model cannot be falsified by our data. Both models combine a preference for being honest with a preference for being seen as honest. In RH+LC, individuals care about lying costs and about the probability of being a liar given their report. In LC-Reputation, individuals care about lying costs and about what an audience observing the report deduces about their lying cost parameter θ. As discussed in Appendix B, the LC-Reputation model predicts equilibrium behavior that is very similar to the General RH+LC model.

All other models fail at least one of the five tests. Looking at Table 1, one can discern certain patterns. The LC model, which is most widely used in the literature, fails two tests, predicting f-invariance and o-invariance. The standard model and its variants fail the same tests. The RH model and almost all other reputation models are unable to explain that all reports are made with positive probability. The CLC model, which is our preferred way to model the effect of descriptive norms, fails three tests, predicting drawing out (when the equilibrium is unique), affinity and o-invariance. All other social comparisons models also predict affinity and o-invariance. Moreover, the LC and the RH models fail the additional tests we implement for settings with more than two states.[36]

We find no significant effect of a change in beliefs, i.e., g-invariance. Even though the point estimate is quite close to zero, this also means that we cannot reject (small) positive or negative effects of a change in beliefs. This may raise the concern that our rejection of many models, in particular the social comparisons models, which all predict affinity, is driven by a lack of power. However, social comparisons models will typically predict quite large responses to shifts in $\hat{G}$. For example, a simple, calibrated version of the CLC model implies that 21 percent of subjects should increase their reports across our $\hat{G}$ treatments, while we find that (in net) 6 percent of subjects *decrease* their report.[37] Regardless of whether our $\hat{G}$ treatments

[35] Figure 2 shows that virtually all treatments with uniform F and 2 or 6 states have full support. This result also holds for all treatments with other, e.g., non-uniform, distributions surveyed in the meta study.

[36] Some researchers, such as Crawford et al. (2013) and Wang et al. (2010), have interpreted evidence from communication games as supportive of level-k reasoning. Our data is consistent with that conclusion. However, our results indicate that level-k reasoning alone is not sufficient to generate the observed behavior. There must be some additional factor, such as LC type motivations. Similarly, level-k reasoning is not necessary to for a model to be consistent with our data.

[37] The 95 percent confidence interval of the difference between the share of high reports across our $\hat{G}$ treatments is from 0.049 to -0.165. We focus on the CLC model as it provides a baseline utility function for modeling social comparisons and cleanly demonstrates the fact that we should expect to see large shifts in our

have enough power or not, even if we interpreted our data on this test as inconclusive, only the Audit model (Appendix B.17) would be added to the set of non-falsified models. Like RH+LC and LC-Reputation, the Audit model can be understood as a combination of a preference for being honest with a preference for being seen as honest.

Importantly, non-uniqueness of equilibria does not affect our overall falsification. Recall that the first and third prediction might not hold when there is more than one equilibrium. All those models that fail the first or third predictions and could feature multiple equilibria also fail additional predictions. Similarly, the models that our data cannot falsify are consistent with the data when the equilibrium is unique.

To show that the surviving models can reproduce the stylized facts of the meta study, we suggest a simple functional form from the class of (separable) RH+LC models. Here, unlike before, we will consider what happens with more than two states. Utility is:

$$\phi(r, c(r,\omega), \Lambda(r); \kappa_1, \kappa_2, \theta_1, \theta_2, r_{max} - r_{min}) = r - \kappa_1\theta_1|r-\omega| - \kappa_2\theta_2\Lambda(r)\frac{(r_{max}-r_{min})}{2} \quad (1)$$

As before, r is the report, ω the true state and $\Lambda(r)$ the fraction of liars at r. κ_1 and κ_2 are the weights of lying costs and reputational costs respectively and are assumed to be fixed across all individuals. θ_1 and θ_2 are the individual-specific weights on the lying cost and the reputation cost and are drawn from a uniform joint distribution on $[0,1] \times [0,1]$. The θs thus serve to create heterogeneity between subjects, while the κs determine the average weight put on the different utility components. r_{min} and r_{max} are the payoffs of the lowest and highest possible report.[38] The utility function is additively separable in the value of the monetary payoff (first term, assuming linear utility in money), the lying cost (second term) and the reputational cost of being thought to lie (third term). The lying cost is linear in the distance between report and true state. In concurrent work, Khalmetski and Sliwka (2016) discuss a similar utility function, combining a fixed cost of lying with concerns for reputation.

This utility function is a extension of that discussed in Section 2.5 and, as the following observation makes clear, implies that reporting behavior is invariant to affine shifts in the monetary payoffs. The proof is in Appendix C.6.

$\hat{G}$ treatments. For details of this calibration see Appendix C.5.

[38] To keep notation simple we assumed fixed payoffs in Section 2 and Appendix B. To match the stylized fact that increasing incentives do not affect reports, we need to add a normalization factor to the reputation cost.

Observation 1 *The set of equilibria implied by the utility function (1) is invariant to affine shifts in payoffs, i.e., if the possible states ω and reports r are replaced by $\alpha\omega + \beta$ and $\alpha r + \beta$, with $\alpha, \beta \in \mathbb{R}$ and $\alpha > 0$.*

We now turn to demonstrating, via a calibration, that this model can match the stylized facts of the meta study. Recall that the stylized facts are: 1) Subjects obtain only about a quarter of potential gains from misreporting. 2) Subjects continue to refrain from lying maximally when stakes are increased. Indeed, the average report does not change when stakes are increased. 3) All possible reports are made with positive probability. 4) Reports that lead to a higher payoff are given more often. 5) Some non-maximal-payoff reports are made more often than their true likelihood.

In order to understand how our model can match the stylized facts, we consider a setting with a uniform distribution F over three possible states, -1, 0 and 1. We use three reports as it is the simplest environment that allows us to study what happens at reports that are neither maximal nor minimal.

If $\kappa_1 = 3$ and $\kappa_2 = 4$, then the utility function (1) can explain all stylized facts of the meta study. The details of the calibration are in Appendix C.6. These parameter values mean that the average individual suffers the equivalent of a 1.5-unit monetary loss due to direct lying costs if they lie so as to increase their "normalized" monetary payoff by one, and that the average individual experiences the equivalent of a 2-unit monetary loss if the probability they are a liar (conditional on their report and the equilibrium strategies of others) goes from 0 to 1. We focus on potential equilibria for the trinary states F where there is no lying downwards. The calibrated model generates a unique equilibrium within this set, which has the following likelihoods for the three reports: $g(-1) = 0.241$, $g(0) = 0.292$, $g(1) = 0.467$. This is very much in line with the results of the meta study. First, the average standardized report turns out to be 0.226, very similar to the value of 0.216 in the meta study. Second, as our observation shows, the reporting distribution is invariant to changes in the incentives. Third, all reports are made with positive probability. In fact, the distribution of reports is very similar to the distributions of reports for uniform distributions shown in Figure A.5, with, in line with the fourth stylized fact, an increasing histogram of reports. Fifth, although in our calibration 0 is reported slightly less often than it is drawn as a state (0.292 versus 0.333) this is actually

in line with what the meta study would predict for three states (in particular Figure A.5). Moreover, for this parameterization, no individual lies downwards.

This parameterization can also accommodate the qualitative features of our experimental treatments quite well. With two states, the model predicts a drawing-in effect moving from F_LOW to F_HIGH, and aversion, affinity or g-invariance (depending on how individuals update their belief about the joint distribution of the θs using $\hat{G}$), and very little upwards lying when the true state is observable, i.e., the same patterns we observe in the data.

5 Conclusion

Our paper attempts to understand the constituent mechanisms that drive lying aversion. Drawing on the extensive experimental literature following the FFH paradigm, we establish some "stylized" facts within the literature, demonstrating that even in one-shot anonymous interactions with experimenters, many subjects do not lie maximally. Our new experimental results, combined with our theoretical predictions, demonstrate that the best description of why people lie is likely to be a model which incorporates multiple types of costs, which have usually been thought of in isolation. In particular, we show that among the many models we consider only a model that combines a desire to be honest with a desire to appear honest cannot be falsified by our data. While we focus on a situation of individual decision making, the utility functions we consider should be present in all situations that involve the reporting of private information, e.g., sender-receiver games, and would there form the basis for the strategic interaction.[39]

What lessons can we draw for policy? The size and robustness of the effect we document suggest that mechanisms that rely on voluntary truth-telling by some participants could be very successful. They could be easier or cheaper to implement and they could achieve outcomes that are impossible to achieve if incentive compatibility is required. Moreover, if the social planner wants to increase truth-telling in the population, our preferred model suggests that lying costs and concerns for reputation are important. Thus, whatever created the lying costs

[39] Focusing more narrowly on experiments, our insights also do not just pertain to setups similar to Fischbacher and Föllmi-Heusi (2013). The matrix task of Mazar et al. (2008), described in the introduction, and other real-effort reporting tasks add ambiguity about the true proportion of correct answers in the population but once our models are adjusted to take the ambiguity into account, they can be directly applied to the Mazar et al. (2008) setting.

in the first place, e.g., education or a Hippocratic oath-type professional norm, is effective and should be strengthened. In addition, one should try to make it harder to lie while keeping a good reputation, e.g., via transparency, naming-and-shaming or reputation systems.

There are at least four potential caveats for these policy implications. First, we wouldn't normally base recommendations on a single lab experiment. Given that our meta study provides very strong, large-scale evidence, however, we feel confident that truth-telling is a robust phenomenon. Second, lab experiments are not ideal to pin down the precise value of policy-relevant parameters. We would thus not put much emphasis on the exact value of, say, the average amount of lying, which we measure as 0.216. However, it is clear that whatever the exact value is, it is far away from 1. Thirdly, none of our results suggests that all people in all circumstances will shy away from lying maximally. Any mechanism that relies on voluntary truth-telling will need to be robust to some participants acting rationally and robust to self-selection of rational participants into the mechanism. Finally, the FFH paradigm does not capture several aspects that could affect reporting. Subjects have to report and have to report a single number. This excludes lies by omission or vagueness (Serra-Garcia et al. 2011). From the viewpoint of the subject, there is also little ambiguity about whether they lied or not. In reality a narrative for reporting a higher state while still maintaining a self-image of honesty might be easier to generate (Falk and Tirole 2016, Mazar et al. 2008).

6 References

J. Abeler. A reporting experiment with Chinese professionals. *mimeo*, 2015.

J. Abeler and D. Nosenzo. Lying and other preferences. *mimeo*, 2015.

J. Abeler, A. Falk, L. Goette, and D. Huffman. Reference points and effort provision. *American Economic Review*, 101(2):470–492, 2011.

J. Abeler, A. Becker, and A. Falk. Representative evidence on lying costs. *Journal of Public Economics*, 113:96–104, 2014.

J. Abeler, D. Nosenzo, and C. Raymond. Preferences for truth-telling. *mimeo*, 2016.

G. Akerlof. Loyalty filters. *American Economic Review*, 73(1):54–63, 1983.

M. Allingham and A. Sandmo. Income tax evasion: A theoretical analysis. *Journal of Public Economic*, 1:323–338, 1972.

J. Andreoni and D. Bernheim. Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636, 2009.

M. Antony, H. Gerhardt, and A. Falk. The impact of food and water deprivation on economic decision making. *mimeo*, 2016.

Y. Arbel, R. Bar-El, E. Siniver, and Y. Tobol. Roll a die and tell a lie–what affects honesty? *Journal of Economic Behavior & Organization*, 107:153–172, 2014.

D. Ariely, X. Garcia-Rada, L. Hornuf, and H. Mann. The (true) legacy of two really existing economic systems. *University of Munich Discussion Paper*, 2014.

G. Aydogan, A. Jobst, F. Loy, N. Müller, and M. Kocher. Oxytocin promotes dishonesty under competition. *mimeo*, 2015.

S. Barfort, N. Harmon, F. Hjorth, and A. L. Olsen. Dishonesty and selection into public service in Denmark: Who runs the world's least corrupt public service? *University of Copenhagen Discussion Paper*, 2015.

D. Batson, D. Kobrynowicz, J. Dinnerstein, H. Kampf, and A. Wilson. In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, 72(6): 1335, 1997.

D. Batson, E. Thompson, G. Seuferling, H. Whitney, and J. Strongman. Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, 77(3):525, 1999.

P. Battigalli and M. Dufwenberg. Guilt in games. *American Economic Review*, 97(2):170–176, 2007.

P. Battigalli and M. Dufwenberg. Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35, 2009.

T. Beck, C. Bühren, B. Frank, and E. Khachatryan. Lying in the face of monitoring, reciprocity, and commitment. *mimeo*, 2016.

R. Bénabou and J. Tirole. Incentives and prosocial behavior. *American Economic Review*, 96 (5):1652–1678, 2006.

M. Blanco and J.-C. Cárdenas. Honesty after a labor relationship. *Universidad del Rosario Discussion Paper*, (2015-37), 2015.

A. Boegli, B. Rockenbach, J. Sobel, and A. Wagner. Why do people tell the truth? A classifcation approach. *mimeo*, 2016.

G. Bolton and A. Ockenfels. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193, 2000.

S. Braun and L. Hornuf. Leadership and persistency in spontaneous dishonesty. *IAAEU Discussion Paper*, 2015.

C. Bryan, G. Adams, and B. Monin. When cheating would make you a cheater: Implicating the self prevents unethical behavior. *Journal of Experimental Psychology: General*, 142 (4):1001, 2013.

A. Bucciol and M. Piovesan. Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology*, 32(1):73–78, 2011.

B. Cadsby, N. Du, and F. Song. In-group favoritism and moral decision-making. *Journal of Economic Behavior and Organization*, 128:59–71, 2016.

J. C. Cardenas and J. Carpenter. Behavioural development economics: Lessons from field labs in the developing world. *Journal of Development Studies*, 44(3):311–338, 2008.

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

J. Chytilova and V. Korbel. Individual and group cheating behavior: a field experiment with adolescents. *IES Working Paper*, 2014.

S. Clot, G. Grolleau, and L. Ibanez. Smug alert! Exploring self-licensing behavior in a cheating game. *Economics Letters*, 123(2):191–194, 2014.

A. Cohn and M. A. Maréchal. Laboratory measure of cheating predicts school misconduct. *CESifo Discussion Paper*, 2015.

A. Cohn, E. Fehr, and M. A. Maréchal. Business culture and dishonesty in the banking industry. *Nature*, 516(7529):86–89, 2014.

A. Cohn, M. A. Maréchal, and T. Noll. Bad boys: How criminal identity salience affects rule violation. *Review of Economic Studies*, 82(4):1289–1308, 2015.

D. Cojoc and A. Stoian. Dishonesty and charitable behavior. *Experimental Economics*, 17 (4):717–732, 2014.

J. Conrads and S. Lotz. The effect of communication channels on dishonest behavior. *Journal of Behavioral and Experimental Economics*, 58:88–93, 2015.

J. Conrads, B. Irlenbusch, R. M. Rilke, and G. Walkowitz. Lying and team incentives. *Journal of Economic Psychology*, 34:1–7, 2013.

J. Conrads, B. Irlenbusch, R. M. Rilke, A. Schielke, and G. Walkowitz. Honesty in tournaments. *Economics Letters*, 123(1):90–93, 2014.

V. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

V. P. Crawford, M. A. Costa-Gomes, and N. Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62, 2013.

G. d'Adda, D. Darai, and R. A. Weber. Do leaders affect ethical conduct? *CESifo Discussion*

Paper, 2014.

Z. Dai, F. Galeotti, and M. C. Villeval. Cheating in the lab predicts fraud in the field. An experiment in public transportations. *Management Science*, forthcoming.

S. Dato and P. Nieken. Compensation and honesty: Gender differences in lying. *mimeo*, 2015.

S. DellaVigna, J. List, U. Malmendier, and G. Rao. Voting to tell others. *NBER Discussion Paper*, 2014.

S. Demichelis and J. W. Weibull. Language, meaning, and games: A model of communication, coordination, and evolution. *American Economic Review*, 98(4):1292–1311, 2008.

S. Di Falco, B. Magdalou, D. Masclet, M. C. Villeval, and M. Willinger. Can transparency of information reduce embezzlement? Experimental evidence from Tanzania. *mimeo*, 2016.

A. Dieckmann, V. Grimm, M. Unfried, V. Utikal, and L. Valmasoni. On trust in honesty and volunteering among Europeans: Cross-country evidence on perceptions and behavior. *European Economic Review*, forthcoming.

A. Diekmann, W. Przepiorka, and H. Rauhut. Lifting the veil of ignorance: An experiment on the contagiousness of norm violations. *Rationality and Society*, 27(3):309–333, 2015.

B. M. Djawadi and R. Fahr. "... and they are really lying": Clean evidence on the pervasiveness of cheating in professional contexts from a field experiment. *Journal of Economic Psychology*, 48:48–59, 2015.

M. Drupp, M. Khadjavi, and M. Quaas. Truth-telling to the regulator? Evidence from a field experiment with commercial fishermen. *mimeo*, 2016.

D. Effron, C. Bryan, and K. Murnighan. Cheating at the end to avoid regret. *Journal of personality and social psychology*, 109(3):395, 2015.

T. Ellingsen and M. Johannesson. Promises, threats and fairness. *The Economic Journal*, 114(495):397–420, 2004.

T. Ellingsen and M. Johannesson. Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008, 2008.

T. Ellingsen and R. Östling. When does communication improve coordination? *American Economic Review*, 100(4):1695–1724, 2010.

C. Engel. Dictator games: A meta study. *Experimental Economics*, 14(4):583–610, 2011.

E. Eyster and M. Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, 2005.

A. Falk and J. Tirole. Narratives, imperatives and moral reasoning. *mimeo*, 2016.

E. Fehr and K. M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly journal of Economics*, 114:817–868, 1999.

U. Fischbacher. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178, 2007.

U. Fischbacher and F. Föllmi-Heusi. Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547, 2013.

A. Foerster, R. Pfister, C. Schmidts, D. Dignath, and W. Kunde. Honesty saves time (and justifications). *Frontiers in Psychology*, 4, 2013.

T. R. Fosgaard. Asymmetric default bias in dishonesty–how defaults work but only when in one's favor. *University of Copenhagen Discussion Paper*, 2013.

T. R. Fosgaard, L. G. Hansen, and M. Piovesan. Separating will from grace: An experiment on conformity and awareness in cheating. *Journal of Economic Behavior & Organization*, 93:279–284, 2013.

A. Frankel and N. Kartik. Muddled information. *mimeo*, 2016.

S. Gächter and J. Schulz. Lying and beliefs. *mimeo*, 2016a.

S. Gächter and J. F. Schulz. Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531:496–499, 2016b.

R. Gibson, C. Tanner, and A. Wagner. Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review*, 103:532–548, 2013.

D. Gill, V. Prowse, and M. Vlassopoulos. Cheating in the workplace: An experimental study of the impact of bonuses and productivity. *Journal of Economic Behavior & Organization*, 96:120–134, 2013.

F. Gino and D. Ariely. The dark side of creativity: original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, 102(3):445, 2012.

U. Gneezy. Deception: The role of consequences. *American Economic Review*, 95(1):384–394, 2005.

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

J. Greene and J. Paxton. Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, 106(30):12506–12511, 2009.

B. Greiner. Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125, 2015.

Z. Grossman. Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization*, 117:26–39, 2015.

R. Halevy, S. Shalvi, and B. Verschuere. Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research*, 40(1):54–72, 2014.

R. Hanna and S.-Y. Wang. Dishonesty and selection into public service. *NBER Discussion Paper*, 2013.

L. Hao and D. Houser. Perceptions, intentions, and cheating. *George Mason University Discussion Paper*, 2013.

D. Harless and C. Camerer. The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–1289, 1994.

B. Herrmann, C. Thöni, and S. Gächter. Antisocial punishment across societies. *Science*, 319 (5868):1362–1367, 2008.

B. Hilbig and C. Hessler. What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology*, 49(2):263–266, 2013.

B. Hilbig and I. Zettler. When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality*, 57:72–88, 2015.

D. Houser, J. A. List, M. Piovesan, A. S. Samek, and J. Winter. On the origins of dishonesty: From parents to children. *European Economic Review*, 82:242–254, 2016.

D. Hruschka, C. Efferson, T. Jiang, A. Falletta-Cowden, S. Sigurdsson, R. McNamara, M. Sands, S. Munira, E. Slingerland, and J. Henrich. Impartial institutions, pathogen stress and the expanding social network. *Human Nature*, 25(4):567–579, 2014.

D. Hugh-Jones. Honesty and beliefs about honesty in 15 countries. *University of East Anglia Discussion Paper*, 2015.

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

C. Jacobsen and M. Piovesan. Tax me if you can: An artefactual field experiment on dishonesty. *Journal of Economic Behavior and Organization*, 124:7–14, 2016.

T. Jiang. Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior & Organization*, 93:328–336, 2013.

T. Jiang. Other-regarding preferences, culture and corruption. *mimeo*, 2015.

N. Johnson and A. Mislin. Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889, 2011.

A. Kajackaite and U. Gneezy. Lying costs and incentives. *UC San Diego Discussion Paper*, 2015.

N. Kartik. Strategic communication with lying costs. *Review of Economic Studies*, 76(4): 1359–1395, 2009.

N. Kartik, M. Ottaviani, and F. Squintani. Credulity, lies, and costly talk. *Journal of Economic Theory*, 134(1):93–116, 2007.

N. Kartik, O. Tercieux, and R. Holden. Simple mechanisms and preferences for honesty. *Games and Economic Behavior*, 83:284–290, 2014b.

K. Khalmetski and D. Sliwka. Disguising lies – image concerns and partial lying in cheating games. *mimeo*, 2016.

B. Kőszegi and M. Rabin. A model of reference-dependent preferences. *Quarterly Journal of Economics*, 121(4):1133–1165, 2006.

B. Kőszegi and M. Rabin. Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073, 2007.

M. Kroher and T. Wolbring. Social control, social learning, and cheating: Evidence from lab and online experiments on dishonesty. *Social Science Research*, 53:311–324, 2015.

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

C.-T. A. Ma and T. McGuire. Optimal health insurance and provider payment. *American Economic Review*, 87(4):685–704, 1997.

H. Mann, X. Garcia-Rada, L. Hornuf, J. Tafurt, and D. Ariely. Cut from the same cloth: Surprisingly honest individuals across cultures. *Journal of Cross-Cultural Psychology*, 47 (6):858–874, 2016.

H. Matsushima. Role of honesty in full implementation. *Journal of Economic Theory*, 139 (1):353–359, 2008.

N. Mazar, O. Amir, and D. Ariely. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6):633–644, 2008.

L. Meub, T. Proeger, T. Schneider, and K. Bizer. The victim matters – Experimental evidence on lying, moral costs and moral cleansing. *cege Discussion Papers*, 2015.

G. Muehlheusser, A. Roider, and N. Wallmeier. Gender differences in honesty: Groups versus individuals. *Economics Letters*, 128:25–29, 2015.

N. Muñoz-Izquierdo, B. Gil-Gómez de Liaño, F. D. Rin-Sánchez, and D. Pascual-Ezama. Economists: cheaters with altruistic instincts. *MPRA Discussion Paper*, 2014.

H. Oosterbeek, R. Sloof, and G. Van De Kuilen. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188, 2004.

D. Pascual-Ezama, T. Fosgaard, J. C. Cardenas, P. Kujal, R. Veszteg, B. G.-G. de Liaño, B. Gunia, D. Weichselbaumer, K. Hilken, A. Antinyan, J. Delnoij, A. Proestakis, M. Tira, Y. Pratomo, T. Jaber-López, and P. Brañas-Garza. Context-dependent cheating: Experimental evidence from 16 countries. *Journal of Economic Behavior & Organization*, 116: 379–386, 2015.

E. Peer, A. Acquisti, and S. Shalvi. 'I cheated, but only a little': Partial confessions to unethical behavior. *Journal of Personality and Social Psychology*, 106(2):202, 2014.

M. Ploner and T. Regner. Self-image and moral balancing: An experimental analysis. *Journal of Economic Behavior & Organization*, 93:374–383, 2013.

K. Popper. *Logik der Forschung*. Julius Springer, Vienna, 1934.

J. Potters and J. Stoop. Do cheaters in the lab also cheat in the field? *European Economic Review*, 87:26–33, 2016.

H. Rauhut. Beliefs about lying and spreading of dishonesty: Undetected lies and their constructive and destructive social dynamics in dice experiments. *PLoS One*, 8(11), 2013.

N. Ruedy and M. Schweitzer. In the moment: The effect of mindfulness on ethical decision making. *Journal of Business Ethics*, 95(1):73–87, 2010.

B. Ruffle and Y. Tobol. Honest on mondays: Honesty and the temporal separation between decisions and payoffs. *European Economic Review*, 65:126–135, 2014.

B. Ruffle and Y. Tobol. Clever enough to tell the truth. *Experimental Economics*, forthcoming.

M. Serra-Garcia, E. Van Damme, and J. Potters. Hiding an inconvenient truth: Lies and vagueness. *Games and Economic Behavior*, 73(1):244–261, 2011.

S. Shalvi and C. De Dreu. Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences*, 111(15):5503–5507, 2014.

S. Shalvi and D. Leiser. Moral firmness. *Journal of Economic Behavior & Organization*, 93: 400–407, 2013.

S. Shalvi, J. Dana, M. Handgraaf, and C. De Dreu. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190, 2011.

S. Shalvi, O. Eldar, and Y. Bereby-Meyer. Honesty requires time (and lack of justifications). *Psychological Science*, 23(10):1264–1270, 2012.

S. Škoda. Effort and cheating behavior: An experiment. *mimeo*, 2013.

D. Stahl and P. Wilson. Experimental evidence on players' models of other players. *Journal*

of economic behavior & organization, 25(3):309–327, 1994.

S. Suri, D. Goldstein, and W. Mason. Honesty in an online labor market. In *Human Computation*, 2011.

S. Tadelis. The power of shame and the rationality of trust. *Haas School of Business Working Paper*, 2011.

I. Thielmann, B. Hilbig, I. Zettler, and M. Moshagen. On measuring the sixth basic personality dimension: A comparison between HEXACO honesty-humility and big six honesty-propriety. *Assessment*, forthcoming.

R. Townsend. Optimal contracts and competitive markets with costly state verification. *Journal of Economic theory*, 21(2):265–293, 1979.

A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

V. Utikal and U. Fischbacher. Disadvantageous lies in individual decisions. *Journal of Economic Behavior & Organization*, 85:108–111, 2013.

C. Vanberg. Why do people keep their promises? An experimental test of two explanations. *Econometrica*, 76(6):1467–1480, 2008.

J. T.-y. Wang, M. Spezio, and C. F. Camerer. Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3):984–1007, 2010.

S. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

B. Waubert De Puiseau and A. Glöckner. Investigating cheating behavior in students compared to the general public. *mimeo*, 2012.

J. Weibull and E. Villa. Crime, punishment and social norms. *SSE/EFI Discussion Paper*, 2005.

O. Weisel and S. Shalvi. The collaborative roots of corruption. *Proceedings of the National Academy of Sciences*, 112(34):10651–10656, 2015.

G. Weizsäcker. Do we follow others when we should? A simple test of rational expectations. *American Economic Review*, 100(5):2340–60, 2010.

M. Wibral, T. Dohmen, D. Klingmüller, B. Weber, and A. Falk. Testosterone administration reduces lying in men. *PLoS One*, 7(10), 2012.

I. Zettler, B. Hilbig, M. Moshagen, and R. de Vries. Dishonest responding or true virtue? A behavioral test of impression management. *Personality and Individual Differences*, 81: 107–111, 2015.

L. Zimerman, S. Shalvi, Y. Bereby-Meyer, et al. Self-reported ethical risk taking tendencies predict actual dishonesty. *Judgment and Decision Making*, 9(1):58–64, 2014.

Online Appendices

A Further results of the meta study

In this appendix, we discuss additional design details and results of the meta study including hypotheses tests. Table A.1 provides descriptive statistics of the independent variables. Figure A.1 marks all countries in which experiments were conducted. The world-wide coverage is quite good, except for Africa and the Middle East.

A.1 Design

We searched in different ways for studies to include in the meta study, using Google Scholar for direct search of all keywords used in the early papers in the literature and to trace who cited those early papers, New Economic Papers (NEP) alerts and emails to professional email lists. We include all studies using the FFH paradigm, i.e., in which subjects conduct a random draw and then report their outcome of the draw, i.e., their state. This excludes sender-receiver games as studied in Gneezy (2005) and the many subsequent papers which use this paradigm or promise games as in Charness and Dufwenberg (2006). We require that the true state is unknown to the experimenter but that the experimenter knows the distribution of the random draw. The first requirement excludes studies in which the experimenter assigns the state to the subjects (e.g., Gibson et al. 2013) or learns the state (e.g., Gneezy et al. 2013). The second requirement excludes the many papers which use the matrix task introduced by Mazar et al. (2008) and comparable real-effort reporting tasks, e.g., Ruedy and Schweitzer (2010). We do include studies in which subjects report whether their prediction of a random draw was correct or not (as in Jiang 2013). Moreover, we require that the payoff from reporting is independent of the actions of other subjects. This excludes games like Conrads et al. (2013) or d'Adda et al. (2014). We do allow that reporting has an effect on other subjects. We need to know the expected payoff level, i.e., the nominal reward and the likelihood that a subject actually receives this nominal reward. If the payoff is non-monetary, we translate the payoff as accurately as possible into a monetary equivalent. We further require that the expected payoff level is not constant, in particular not always zero, i.e., making different reports has to lead to different consequences. We exclude studies in which subjects could self-

select into the reporting experiment after learning about the rules of the experiment. This excludes the earliest examples of this class of experiments, Batson et al. (1997) and Batson et al. (1999). Finally, we exclude random draws with non-symmetric distributions, except if the draw has only two potential states. We exclude such distributions since the average report for asymmetric distributions with many states is difficult to compare to the average report of symmetric distributions. This only excludes Cojoc and Stoian (2014) and two of our treatments reported in this paper.[40]

A.2 Influence of treatment variables

In this section, we further explore the effect of variables that differ between treatments and test the statistical significance of those effects. For such treatment-level variables, we use two complementary identification strategies. First, we can assume that the error term is independent of the explanatory variables once we control for all observable variables. This conditional-independence assumptions allows us to interpret the regression coefficients as the causal effects of the explanatory variables. While the conditional-independence assumption is usually regarded as a quite strong assumption, it is less strong in our setting for several reasons. Economics laboratory experiments are highly standardized and lab experiments are run with very abstract framing, usually eschewing any context and just describing the rules of the games. Both of these arguments mean that the importance of omitted variables is likely to be limited. Moreover, researchers usually select the design of their experiments with regard to the research question they are interested in and not with regard to characteristics of the local subject pool. Reverse causality is thus also unlikely. Results are reported in Table A.2, columns 1 and 2. We include all explanatory variables that vary across more than one treatment.[41]

[40] We adjust the distribution of standardized reports of experiments with asymmetric distributions appropriately to be in line with the intuitions regarding the average standardized report discussed in Section 1.

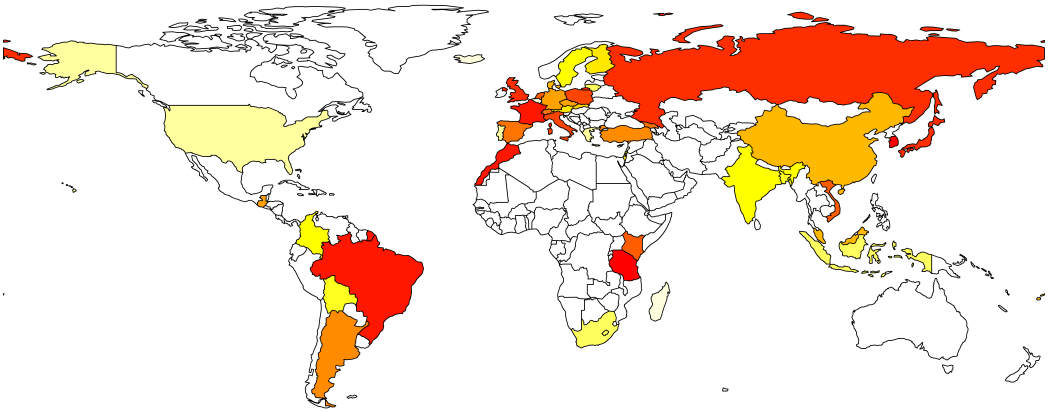
[41] We restrict explanatory variables in this way since otherwise any treatment fixed effect could be an explanatory variable. Given that we include 362 treatments this would become unwieldy.

Table A.1: Meta study: descriptive statistics

	Mean	# Subjects
Treatment-level variables		
Maximal payoff from misreporting (in 2015 USD)	4.677	32503
1 if student subjects	0.602	32503
1 if repeated	0.319	32503
1 if online/telephone	0.256	32503
1 if control rolls suggested	0.230	32503
1 if reporting about state of mind	0.165	32503
1 if info about behaviour of other subjects available	0.009	32503
1 if report reduces payoff of another subject	0.021	32503
Year experiment conducted	2012.755	32503
Author affiliation		
1 if economics	0.697	32503
1 if psychology	0.263	32503
1 if sociology	0.040	32503
Method of randomization		
1 if coin toss	0.399	32503
1 if die roll	0.539	32503
1 if draw from urn	0.061	32503
True distribution		
1 if two outcomes non-uniform	0.147	32503
1 if two outcomes uniform	0.446	32503
1 if other uniform	0.341	32503
1 if bell shaped	0.066	32503
Individual-level variables		
1 if female	0.475	17508
Age	26.158	12052
Field of study		
1 if economics/management student	0.239	5012
1 if psychology student	0.028	5012
1 if other student	0.733	5012
# Decisions	108140	
# Subjects	32503	
# Treatments	362	
# Studies	72	

Notes: The means are computed on subject level. The maximal payoff refers to the maximal nominal payoff times the probability a subject is actually paid and is converted using PPP.

Figure A.1: Average report by country



Notes: The figure depicts the average standardized report per country. The darker the color, the higher the average report. For exact country averages see Figure A.4.

The second identification strategy we employ makes use of the random assignment of subjects to treatments within study (and the few within-subject experiments). As long as we control for study fixed effects and as long as treatments within a study only differ along one dimension, this eliminates all omitted variables. This is thus a very clean form of identification. The drawback is that the sample is reduced to only those studies that vary the parameter of interest. Potentially, no such study exists. We thus report results of regressions using both identification strategies, the specifications with study fixed effects are in Table A.2, columns 3 to 9. In the regressions we cluster standard errors on each subject, thus treating repeated decisions by the same subject as dependent but treating the decisions by different subjects as independent. This is the usual assumption for experiments that study individual decision making. This assumption is also made in basically all studies we include in the meta study.[42] In the regressions relying on conditional independence, we also report a specification which clusters on study to allow for dependencies within study. Independent of clustering, we weight one decision as one observation in all regressions.[43]

Incentive level: Figure 1 showed that the level of incentives has only a very small effect on the standardized report. The corresponding regressions are in Table A.2, columns

[42] In two studies, Diekmann et al. (2015) and Rauhut (2013), subjects are shown the reports of other subjects in their matching group before making a decision. For these studies we cluster on matching group rather than on individual.

[43] If we weight by subject, results are broadly similar. The overall average standardized report, for example, is then 0.299 instead of 0.216.

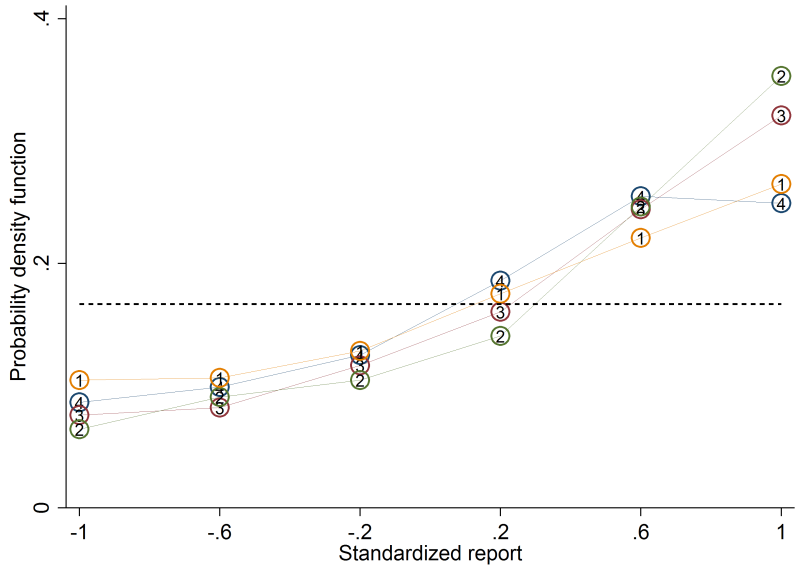
Table A.2: Regressions of treatment-level variables

Dependent variable: Standardized report									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Maximal payoff from misreporting	-0.002***	-0.002	0.003***						
	(0.001)	(0.004)	(0.001)						
1 if student subjects	0.121***	0.121***		0.100***					
	(0.011)	(0.039)		(0.026)					
1 if repeated	-0.046***	-0.046							
	(0.014)	(0.053)							
1 if online/telephone	-0.000	-0.000			0.027				
	(0.012)	(0.053)			(0.032)				
1 if control rolls suggested	-0.062***	-0.062				0.168***			
	(0.019)	(0.062)				(0.052)			
1 if reporting about state of mind	-0.035**	-0.035					0.160***		
	(0.016)	(0.062)					(0.034)		
1 if info about behaviour of other subjects available	0.045	0.045***						0.046	
	(0.045)	(0.009)						(0.049)	
1 if report reduces payoff of another subject	0.010	0.010							-0.119***
	(0.029)	(0.098)							(0.038)
Year experiment conducted	-0.011***	-0.011*							
	(0.003)	(0.006)							
Additional controls	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Author affiliation FE	Yes	Yes	No	No	No	No	No	No	No
Randomization method FE	Yes	Yes	No	No	No	No	No	No	No
True distribution FE	Yes	Yes	No	No	No	No	No	No	No
Country FE	Yes	Yes	Yes	Yes	No	No	No	No	No
Study FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# Decisions	108140	108140	9756	9859	1214	577	1200	12660	2170
# Subjects	32503	32503	7718	4792	906	577	1200	1380	2104
# Treatments	362	362	92	40	12	14	16	18	22
# Studies	72	72	10	8	3	3	1	4	4
# Clusters	32476	72	7718	4792	906	577	1200	1203	2104

Notes: OLS regressions. Robust standard errors clustered on individual subjects (on studies in column 2) are in parentheses. The sample in columns 3 to 7 is restricted to those studies in which the independent variable of interest varies. Maximal payoff from misreporting is the difference of the highest and lowest potential payoff (converted by PPP to 2015 USD). The fixed effects control for author affiliation (economics, psychology, sociology), randomization method (die roll, coin toss, draw from urn), true distribution (asymmetric with two outcomes, uniform with two outcomes, other uniform, bell shaped), country and/or study. Significance at the 1, 5, and 10 percent level is denoted by ***, **, and *, respectively.

1 and 2. An increase of the potential payoff by 1 USD changes the standardized report by -0.002. In column 3, we only use within-study variation for identification. We restrict the sample to those studies which vary the payoff level between treatments. A couple of studies vary payoff level and another variable independently. In the regression, we control for those other variables and mark this as "Additional controls: Yes" in the table. If we cannot properly control for within-study variation, we exclude the affected treatments (we do the same in columns 4–9). The resulting coefficient of 0.003 is very similar to the coefficient derived under the conditional-independence assumption. Even though the coefficients are very small, given our large sample size, both are significantly different from zero. Taken together, this provides converging evidence that the average amount of lying does not change much if stakes are increased. This result is further corroborated by Figure A.2. This figure shows the distribution of reports for experiments using a uniform distribution with six states. We collapse treatments by the potential payoff from misreporting and show the distributions for the four quartiles (weighted by number of subjects). The line marked by "1" is the distribution of the treatments with the lowest payoffs while the line marked "4" represents the treatments with the highest payoffs. Overall, distributions do not differ systematically by payoff level. All distributions are increasing and the second highest state is always reported with more than 1/6 probability. Overall, neither the average report nor the reporting pattern is affected by the payoff level.

Figure A.2: Distribution of reports by incentive level



Notes: The figure depicts the distribution of reports for treatments that use a uniform distribution with six states and linear payoff increases. Treatments are collapsed into quartiles by the level of the maximal payoff from misreporting. The line marked by "1" is the distribution of the treatments with the lowest payoffs while the line marked "4" represents the treatments with the highest payoffs. The dashed line indicates the truthful distribution at 1/6.

Repetition: We next analyze the effect of repetition to ascertain whether subjects converge to the standard economic predictions once they learn and gain experience.

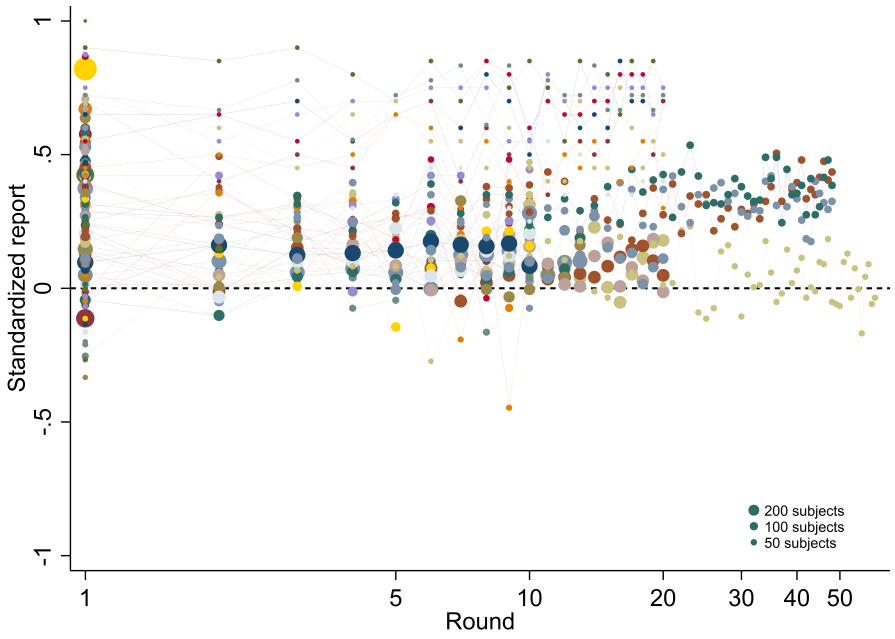
Finding 12 *Reports do not change much under repetition.*

The regressions in Table A.2, columns 1 and 2, show that experiments with repeated reports induce slightly lower reports than one-shot experiments. There are no studies which compare one-shot with repeated implementations directly. We can still use within-study variation to estimate the effect of repetition by comparing reports in early vs. late rounds. Figure A.3 plots the average standardized report by treatment and round. One-shot treatments are shown as round 1. There is no clear trend over rounds. Results of the corresponding regression analysis are reported in Table A.3, column 1. We control for treatment fixed effects and thus restrict the sample to repeated studies, as only they have within-treatment variation in rounds. For those studies, round has a very small, though significantly positive effect. This very weak time trend contrasts strongly with, e.g., public goods games experiments in which

a strong convergence over time to the standard prediction can be observed (e.g., Herrmann et al. 2008).

Taken together, this shows that the overall low reports are robust to learning and experience. Moreover, this corroborates our theoretical approach to model each reporting decision as separate and independent.

Figure A.3: Average standardized report by round



Notes: The figure plots standardized report over the rounds in the experiment. Standardized report is on the y-axis. A value of 0 means that subjects realize as much payoff as a group of subjects who all tell the truth. A value of 1 means that subjects all report the state that yields the highest payoff. The round of the experiment is on the x-axis. One-shot experiments are shown as round 1. Each bubble represents the average standardized report of one treatment in a given round and the size of a bubble is proportional to the number of subjects in that treatment.

Subject pool: Student samples report significantly higher states than samples taken from the general population. Since the latter samples are likely to also include some current students and many subjects who used to be students, these regressions likely underestimate the difference between students and non-students. Students and non-students differ in many respects. We show below that age plays no significant role in the student effect. However, cognitive skills, socio-economic background, current income, etc. could all be part of it.

Reporting channel: While most experiments were conducted in a laboratory, about a third of experiments were conducted remotely via telephone or an online survey. Since the experimenter controls the entire environment of the lab, subjects might fear to be observed, say, by secret cameras. Such an observation is impossible if reports are done by telephone or an online survey since the (physical) random draw is done remotely and thus entirely unobservable. The channel of reporting could also have a direct effect on reporting. We find that reports done remotely do not differ from reports in the lab.

Control rolls suggested: In about one in five experiments the experimenter suggested explicitly that subjects use the randomization device (most often a die) several times in a row. We find mixed evidence for the effect of this suggestion. In column 1, it reduces reports significantly while it increases reports significantly in column 6.

Reporting about state of mind: Following Jiang (2013) and Greene and Paxton (2009), quite a few studies ask subjects to privately make a prediction about the outcome of a random draw. The random draw is usually implemented on a computer and the outcome is known to the experimenter. The report consists of the subject claiming whether their prediction was correct or not. The overall structure is very similar to a standard coin-flip experiment: whether the report is truthful cannot be judged individually by the experimenter, but the experimenter knows the true distribution of states. The only difference is thus whether the subject makes a report about a state of mind or a physical state of the world. The results in columns 1 and 2 show that reporting about a state of mind leads to significantly lower reports. However, the one study which tested this difference directly (Kajackaite and Gneezy 2015) finds that reports about a state of mind are significantly higher (column 7).

Information about others' behavior: In a handful of experiments, subjects were given information about the past behavior of other subjects in similar experiments. This does not affect the average report significantly, except in column 2.

From whom payoff is taken: In most experiments, subjects take money from the experimenter or the laboratory if they report higher states. In a few treatments, subjects' reports instead reduces the payoff of another subject, i.e., the total amount of payoff allocated to two subjects is fixed and the report decides how much of that fixed amount goes to the reporting subject. Columns 1 and 2 indicate that this has no effect on reports while column 9 indicates a significant reduction in reports. Since there are only few studies which implement

this paradigm, or the previous one, (see Table A.1) we would think that the estimates in columns 1 and 2 are less reliable for these two variables.

Year of experiment: Reports have decreased slowly over time but this effect is very small, given that the earliest experiments were conducted in 2005.

Author affiliation: Reports in studies conducted by sociologists, economists or psychologists do not differ significantly.

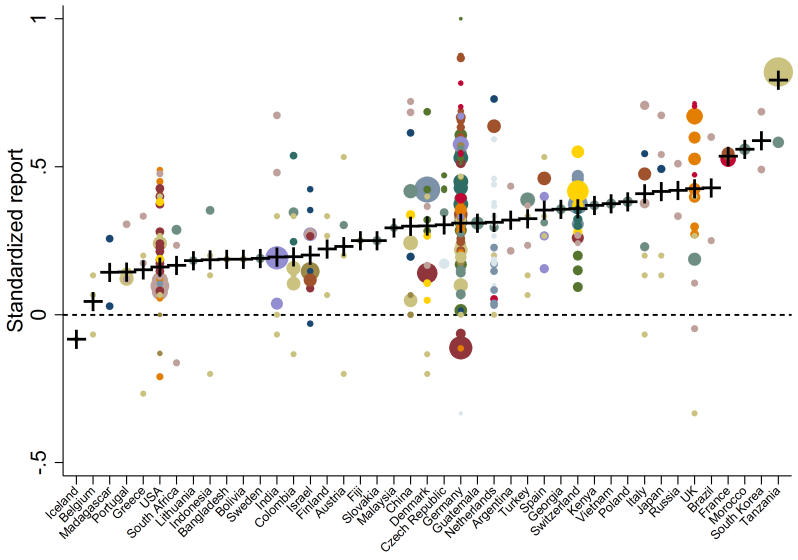
Randomization method: Reports do not differ significantly when a die roll or a coin toss is used. Studies using a draw from an urn yield lower reports.

True distribution: Reports for different uniform distributions do not differ significantly (see also Figure A.5). Compared to uniform distributions, asymmetric distributions have higher reports and bell-shaped distributions have lower reports.

A.3 Further robustness checks

Behavior is surprisingly robust across countries. Figure A.4 plots average standardized reports by country. The country average is marked by a cross. Some of the cross-country variation comes from studies that run the same design across different countries while some of the variation is coming from researchers using convenience samples of subjects in different countries. For those countries for which we have a decent amount of data, the country averages vary only relatively little, from about 0.1 to about 0.4 (except Tanzania). Country fixed effects are only able to explain 0.197 of the between-treatment variation (adjusted $R^2 = 0.088$). For detailed analyses of what drives cross-country differences, see, e.g., Pascual-Ezama et al. (2015), Hugh-Jones (2015), Mann et al. (2016) or Gächter and Schulz (2016b).

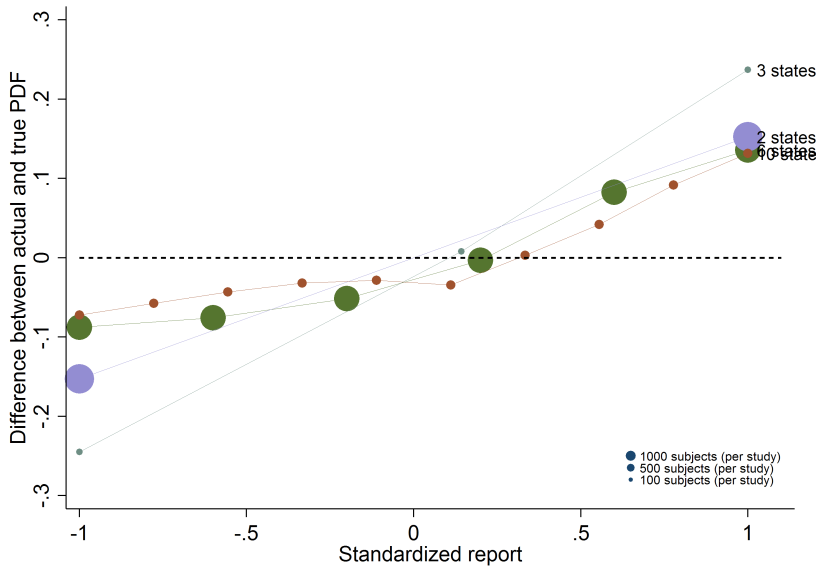
Figure A.4: Average standardized report by country



Notes: The figure plots standardized report against country. Standardized report is on the y-axis. A value of 0 means that subjects realize as much payoff as a group of subjects who all tell the truth. A value of 1 means that subjects all report the state that yields the highest payoff. Each bubble represents the average standardized report of one treatment and the size of a bubble is proportional to the number of subjects in that treatment. The cross is the average per country.

In Figure 2, we showed for uniform distributions with two and six states that the distribution of reports is increasing and has full support. This finding generalizes to uniform distributions with different number of states. Figure A.5 demonstrates that the distribution of reports is actually quite similar for experiments with different numbers of states, only the distribution for the one study using a 3-state distribution is a little steeper.

Figure A.5: Distribution of reports (uniform true distributions)



Notes: The figure depicts the difference between the actual and the truthful distribution of reports for treatments that use a uniform true distribution and linear payoff increases. Treatments are collapsed by the number of states, 2, 3, 6, or 10. The dashed line at 0 indicates the truthful distribution. The size of a bubble is proportional to the number of subjects in the treatments with a given number of states.

A.4 Heterogeneous treatment effects

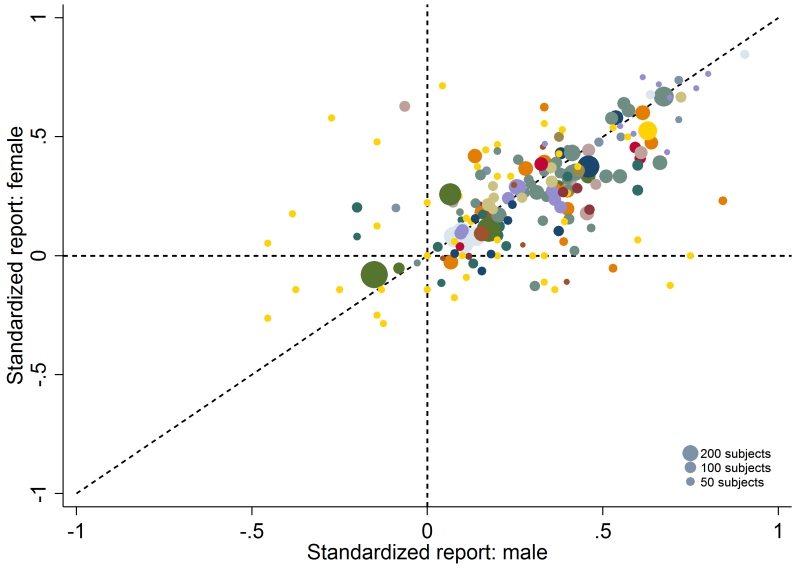
So far, we have focused on variables that differed only on treatment level. For a subset of studies we also have data on individual-level variables, namely gender, age and field of study.

Finding 13 *Women report lower states than men.*

Figure A.6 shows the effect of gender on reports. The majority of treatments is below the 45° line, indicating that female subjects report lower numbers than male subjects. However, there are also many treatments in which women report higher numbers than men. We test the significance of this effect by regressing the report on a gender dummy and controlling for treatment fixed effects. We thus only use within-treatment variation. The results are presented in Table A.3, column 2: women's standardized report is on average 0.057 lower than men's. This effect is highly significant. Figure A.7 shows the distribution by gender of all treatments that use a uniform distribution with six states for which we have gender data.

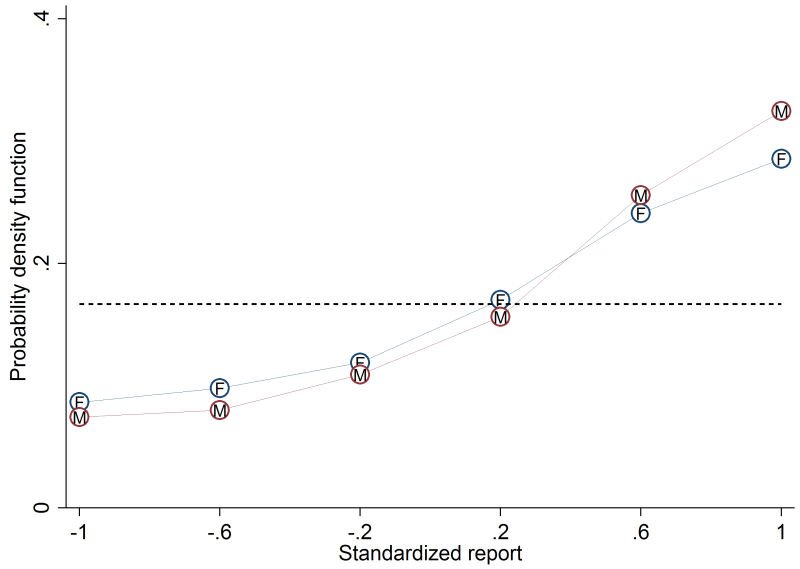
Men are generally less likely to report lower states and more likely to report higher states.

Figure A.6: Average standardized report by gender



Notes: The figure plots the average standardized report of male subjects (x-axis) vs. the average standardized report by female subjects (y-axis). A standardized report of 0 means that subjects realize as much payoff as a group of subjects who all tell the truth. A value of 1 means that subjects all report the state that yields the highest payoff. Data is restricted to those treatments where male and female subjects participated. The size of a bubble is proportional to the number of subjects in that treatment.

Figure A.7: Distribution of reports by gender



Notes: The figure depicts the distribution of reports for treatments that use a uniform distribution with six states and linear payoff increases, collapsed by gender. The line marked "F" is the distribution of female subjects and the line marked "M" is the distribution of male subjects. The dashed line indicates the truthful distribution at 1/6.

Table A.3: Regressions of individual-level variables

Dependent variable: Standardized report

	(1)	(2)	(3)	(4)	(5)
Round	0.001**				
	(0.000)				
1 if female		-0.057***			
		(0.009)			
Age			-0.001	-0.002	
			(0.001)	(0.004)	
Age squared				0.000	
				(0.000)	
1 if economics/management student					-0.016
					(0.023)
1 if psychology student					-0.068
					(0.078)
Treatment FE	Yes	Yes	Yes	Yes	Yes
# Decisions	73582	82395	35675	35675	8063
# Subjects	4862	16736	11319	11319	4383
# Treatments	43	196	120	120	49
# Studies	11	38	27	27	8
# Clusters	4806	16680	11319	11319	4383

Notes: OLS regressions. Robust standard errors clustered on individual subjects are in parentheses. The sample in each specification is restricted to those treatments in which the independent variable(s) vary. Significance at the 1, 5, and 10 percent level is denoted by ***, **, and *, respectively.

Finding 14 *Age and field of study have no effect on reporting behavior.*

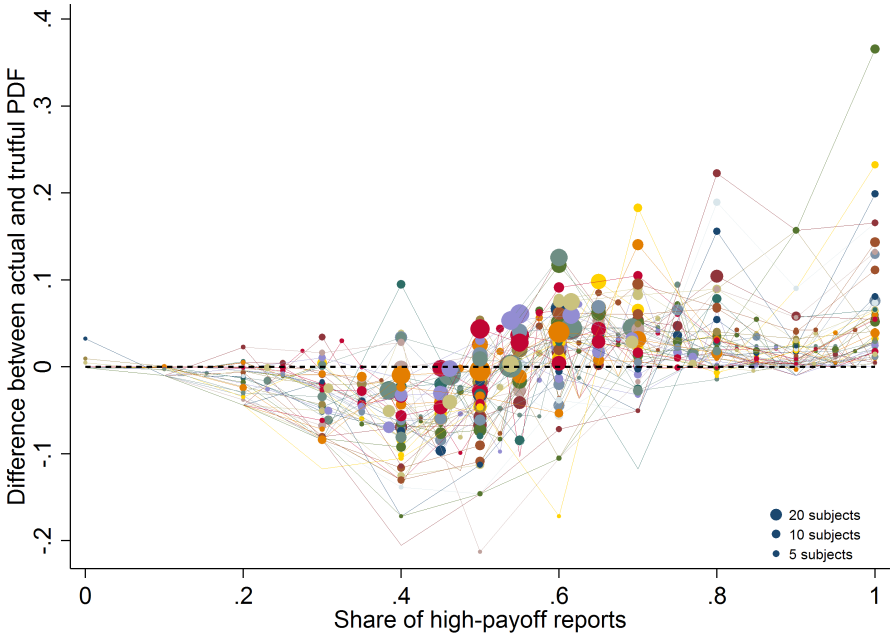
Older subjects tend to report lower numbers but this effect is not significant (Table A.3, columns 3 and 4). While students in general make higher reports than non-students, we do not find an effect of field of study. We focus on comparing economics/management students to psychology students to students studying all other fields. Economics and psychology students report lower states but this effect is not significant (Table A.3, column 5).

Up to here, we have shown that reporting is far from the standard rational prediction of a standardized report of +1 in the entire sample and in all sub-groups defined by observable

characteristics, e.g., gender. However, maybe there is a sub-group, which we cannot identify by observable variables, which does behave according to the standard prediction. For this we would need to identify for each individual whether they lied or not, which is not possible for the one-shot experiments. However, if we aggregate the many reports of an individual subject in repeated experiments, we can test for each individual subject whether their sequence of reports could be generated by truth-telling. In particular, it is increasingly unlikely to repeatedly draw the highest-payoff state. Note that we depart for this analysis from our usual approach of treating each decision as separate and independent. For example, if subjects care about being perceived as truthful, the predicted behavior depends on whether subjects and the audience treat each decision separately or not.

In Figure A.8 we focus on experiments in which subjects repeatedly report the state they drew of a uniform distribution with two states and add up the number of times a subject reported the high-payoff state. To make experiments with different numbers of rounds comparable, we plot the share of the potential high-payoff reports on the x-axis and the difference between the observed distribution and the truthful binomial distribution on the y-axis. Reporting the highest-payoff state in each round is the standard rational prediction. This reporting pattern could have resulted from truth-telling only with a minuscule chance of $1/2^{10}$ to $1/2^{40}$. As one can see in the figure, more subjects always report the high-payoff state than would be expected under full truth-telling. However, the overall share of subjects at this point is surprisingly small. Only 3.6 percent of subjects always report the high-payoff state and only 6.7 percent report it more than 80 percent of the time (the size of the bubbles is proportional to the number of subjects making the respective report). Overall, this suggests that also individually, people are far from the standard prediction.

Figure A.8: Distribution of sum of reports (repeated reports of two-state distributions)



Notes: The figure displays the distribution of the sum of standardized reports in experiments in which subjects repeatedly report the state of a uniform distribution with two states.. Each line represents one treatment. The share of the potential high-payoff reports is on the x-axis. On the y-axis is the difference between the actual and and the truthful PDF. The size of a bubble is proportional to the number of subjects in a given treatment at this share of high-payoff reports.

A.5 List of studies included in the meta study

Table A.4: List of studies included in the meta study

Study	# treatments	# subjects	Country	Randomization method	True distribution
Abeler et al. (2016) *	7	1124	United Kingdom	multiple	multiple
Abeler et al. (2014) *	4	1102	Germany	coin toss	multiple
Abeler (2015) *	1	60	China	draw from urn	1D10
Abeler and Nosenzo (2015) *	9	507	Germany	draw from urn	1D10
Antony et al. (2016) *	2	200	Germany	die roll	1D6
Arbel et al. (2014) *	2	399	Israel	die roll	1D6
Ariely et al. (2014)	1	188	Germany	die roll	1D6
Aydogan et al. (2015)	2	120	Germany	coin toss	2D2
Barfort et al. (2015)	1	862	Denmark	die roll	asy. 1D2
Beck et al. (2016) *	6	128	Germany	die roll	1D6
Blanco and Cárdenas (2015)	2	103	Colombia	die roll	1D6
Braun and Hornuf (2015)	7	342	Germany	die roll	1D2
Bryan et al. (2013) *	3	269	USA	coin toss	1D2
Bucciol and Piovesan (2011) *	2	182	Italy	coin toss	1D2
Cadsby et al. (2016)	1	90	China	die roll	1D6
Chytilova and Korbel (2014) *	1	117	Czech Republic	die roll	1D6
Clot et al. (2014) *	2	98	Madagascar	die roll	1D6
Cohn et al. (2014) *	8	563		coin toss	1D2
Cohn et al. (2015) *	4	375	Switzerland	coin toss	1D2
Cohn and Maréchal (2015)	1	162	Switzerland	coin toss	1D2
Conrads et al. (2013) *	4	554	Germany	die roll	1D6
Conrads and Lotz (2015) *	4	246	Germany	coin toss	4D2
Dai et al. (forthcoming)	2	384	France	die roll	1D3
Dato and Nieken (2015)	1	288	Germany	die roll	1D6
Dieckmann et al. (forthcoming)	5	1015	multiple (5)	coin toss	1D2
Diekmann et al. (2015) *	4	466	Switzerland	die roll	1D6
Di Falco et al. (2016)	1	1080	Tanzania	coin toss	1D2
Djawadi and Fahr (2015)	1	252	Germany	draw from urn	asy. 1D2
Drupp et al. (2016)	4	170	Germany	coin toss	4D2
Effron et al. (2015) *	8	2151	USA	coin toss	1D2
Fischbacher and Föllmi-Heusi (2013) *	5	979	Switzerland	die roll	1D6
Foerster et al. (2013) *	1	28	Germany	die roll	12D8
Fosgaard (2013) *	1	505	Denmark	die roll	2D6
Fosgaard et al. (2013) *	4	209	Denmark	coin toss	1D2
Gächter and Schulz (2016b) *	23	2568	multiple (23)	die roll	1D6
Gino and Ariely (2012)	8	304	USA	die roll	1D6
Halevy et al. (2014) *	1	51	Netherlands	die roll	1D6
Hanna and Wang (2013)	2	826	India	die roll	1D6
Hilbig and Hessler (2013) *	6	765	Germany	die roll	asy. 1D2
Hilbig and Zettler (2015) *	6	342	Germany	multiple	asy. 1D2

Study	# treatments	# subjects	Country	Randomization method	True distribution
Houser et al. (2016)	3	740	Germany	coin toss	1D2
Houser et al. (2016) *	2	72	USA	coin toss	asy. 1D2
Hruschka et al. (2014)	8	223	multiple (6)	die roll	1D2
Hugh-Jones (2015) *	30	1390	multiple (15)	coin toss	1D2
Jacobsen and Piovesan (2016)	3	148	Denmark	die roll	1D6
Jiang (2013) *	6	39	Netherlands	die roll	1D2
Jiang (2015) *	4	224	multiple (4)	die roll	1D2
Kajackaite and Gneezy (2015)	17	1303	multiple (2)	multiple	multiple
Kroher and Wolbring (2015) *	9	384	Germany	die roll	1D6
Mann et al. (2016)	10	2179	multiple (5)	die roll	1D2
Meub et al. (2015)	2	94	Germany	die roll	1D2
Muehlheusser et al. (2015) *	1	108	Germany	die roll	1D6
Muñoz-Izquierdo et al. (2014) *	3	270	Spain	coin toss	1D2
Pascual-Ezama et al. (2015) *	48	1440	multiple (16)	coin toss	1D2
Ploner and Regner (2013) *	6	316	Germany	die roll	1D2
Potters and Stoop (2016) *	6	102	Netherlands	draw from urn	1D2
Rauhut (2013) *	3	240	Switzerland	die roll	1D6
Ruffle and Tobol (2014) *	1	427	Israel	die roll	1D6
Ruffle and Tobol (forthcoming) *	1	156	Israel	die roll	1D6
Shalvi et al. (2011) *	2	129	USA	die roll	1D6
Shalvi et al. (2012) *	4	144	Israel	die roll	1D6
Shalvi and Leiser (2013) *	2	126	Israel	die roll	1D6
Shalvi and De Dreu (2014) *	8	120	Netherlands	coin toss	1D2
Suri et al. (2011)	3	674	multiple (2)	die roll	multiple
Thielmann et al. (forthcoming) *	1	152	Germany	coin toss	asy. 1D2
Utikal and Fischbacher (2013)	2	31	Germany	die roll	1D6
Škoda (2013)	3	90	Czech Republic	die roll	1D6
Waubert De Puiseau and Glöckner (2012)	4	416	Germany	coin toss	5D2
Weisel and Shalvi (2015) *	9	178	multiple (2)	die roll	asy. 1D2
Wibral et al. (2012)	2	91	Germany	die roll	1D6
Zettler et al. (2015) *	1	134	Germany	coin toss	asy. 1D2
Zimerman et al. (2014) *	1	189	Israel	coin toss	1D2

Notes: Studies for which we obtained the full raw data are marked by *. 1DX refers to a uniform distribution with X outcomes. A coin flip would thus be 1D2. ND2 refers to the distribution of the sum of N uniform random draws with two outcomes. Asymmetric 1D2 refers to distributions with two outcomes for which the two outcomes are not equally likely.

B Additional Models

In this section we discuss the remaining models listed in Table 1. After describing each model, we present either informally (for those that are relatively straightforward) or formally the predictions of each model in each of our five dimensions. Many of the models build on the intuitions of the models discussed in the body of the paper. Similarly the proofs regarding the predictions of the models rely on similar mechanisms. Thus, we will often refer readers to the results of Appendix C.

B.1 Inequality Aversion

This model captures the widely discussed notion that individuals care about how their monetary payoff compares to the payoff of others as in, e.g, Fehr and Schmidt (1999) or Bolton and Ockenfels (2000). In our formal model we will build off the intuition of the latter, although similar results hold for a model in line with the former. We suppose that individuals care not just about their own payoff, but also the average payoff. Formally, utility is

$$\phi(r, \nu\varsigma(r - \bar{r}))$$

where $\bar{r}$ is the mean report. ς is a function that maps the difference between an individual's payoff and the average payoff to a utility cost. It has a minimum at 0 and is increasing in the absolute distance from 0. ν is the individual-specific parameter that governs the weight that an individual applies to social comparisons which we suppose has a non-atomic distribution $\widetilde{H}$ on $[0, 1]$. Thus, it is the social preferences analog of θ. We suppose that ϕ is strictly increasing in its first argument and strictly decreasing it its second, i.e., individuals like money and dislike inequality. Moreover, there may be multiple equilibria. For example, if all individuals face a sufficiently strong cost of deviation from the mean report, then for any report r, everyone reporting r is an equilibrium. An equilibrium will exist because of the continuity of ϕ and ς and the property that $\bar{r}$ is continuous in the threshold types (where the threshold is in ν).

Observation 2 *Suppose individuals have Inequality Aversion utility. The set of equilibria does not change with F, and we observe affinity. The observability of the state does not affect the set of equilibria. Equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: Observe that utility does not depend directly on the drawn state ω. This implies that the distribution of reports does not depend on F and so the set of equilibria will not change with F. Although there may be multiple equilibria, because G enters in the the utility function directly (because G has a one-to-one mapping with $\bar{r}$) we can still make predictions regarding the effect of $\hat{G}$. We will observe affinity since ς has a minimum when $r = \bar{r}$. The distribution of reports will not depend on observability of the state since utility does not depend on any inference of others and so the set of equilibria will not change with observability. However, because individuals are concerned about others' reports, they could lie down in equilibrium. Full-support equilibria can exist for any F where $f(\omega_0) \in (0, 1)$ whenever there is a strong enough cost from having a different report from everyone else.□

B.2 Kőszegi-Rabin

Kőszegi and Rabin (2006) suggest a widely used model of expectations dependent reference-dependence. The simplest adaptation of their model to our setting is to assume that individuals have a reference point regarding their payoffs, and have no cost of lying. We denote the equilibrium mapping from state to report as a. In this case, utility would be:

$$\phi(r, \omega, a; \rho, \lambda) = u(r) + \rho \sum_k \lambda^{\mathbb{I}} |u(r) - u(a(\omega_k))| f(\omega_k)$$

An equilibrium a is a mapping such that if a maps $\hat{\omega}$ to $\hat{r}$, then the argmax of $\phi(r, \hat{\omega}, a)$ is $\hat{r}$. ρ is a scalar that represents the weight on gain-loss utility. $\lambda^{\mathbb{I}}$ is an indicator function that takes on a value of 1 if $u(r) - u(a(\omega_k))$ is positive, and $-\lambda$ otherwise. $\lambda > 1$ represents loss aversion. We suppose that the reference point (i.e. the distribution of reports) is established after F is known, but before ω is drawn. Similar to the standard model, in this model all individuals will report the highest possible report, independent of F, $\hat{G}$ or the observability of the report (and so the reporting distribution is unique).

B.3 Generalized Reputation Models

In this section we consider a generalized model of reputation which nests all our other reputation models (except the collective reputation model). We do this to prove a general observation regarding the entire class of models. In this generalized reputation model, individuals do not suffer any costs from mis-reporting in and of itself. Instead, the cost comes from perceived

inferences others make about an individual given their action; i.e. their reputation. Thus, individuals are concerned about what action they take, as well as the equilibrium strategies of others, as this is what is used to infer about unobserved private information (such as whether a given person is lying, or their type). Thus, individuals' utility depends on their action, their type, the equilibrium mapping between types, states, and actions, and the distribution of actions. In order to simplify notation we define an equilibrium mapping Ψ: $\omega \times \theta \to \Delta(r)$, where $\Delta(r)$ is the probability distribution over actions.

Formalizing these intuitions, the utility function is then

$$\phi(r, \Psi, G; \theta)$$

We, analogously to before, suppose that ϕ is increasing in the first element. Whether and how G changes with respect to F and $\hat{G}$ depends on the specifics of the model, as does the existence of an equilibrium with a uniform F. Moreover, whether or not observability matters will depend on the exact structure of the model; we have examples where it does (RH) and examples where it doesn't (reputation for being not greedy).

However regardless of the specifics of the functional form, one implication about this class of models is that if the equilibrium has full support some individuals must lie down when the state is unobservable.

Observation 3 *Under Generalized Reputation Models if the equilibrium has full support, a positive measure of individuals must lie downwards when the state is not observed by outsider.*

Proof: First, fixing an equilibrium, observe that if an individual with state ω and parameter θ weakly prefers r over r', then it must be the case that an individual with state ω' and parameter θ weakly prefers r over r'. To see this, note that utilities for both individuals are independent of their actual draw of ω.

Now suppose an equilibrium has full support. Observe that the expected conditional (on ω) distribution of θ is independent of ω. Now, observe that conditional on drawing a particular state, individuals will follow a threshold rule — people with $\theta \geq \bar{\theta}$ will give one report, and everyone else a different report.

Without loss of generality, suppose that this implies that a positive measure of individuals

with state ω_0 report r_0. This means that there exists a set of θ's with positive measure that strictly prefer reporting r_0 conditional on drawing ω_0. Thus the exact same set of θ's strictly prefer reporting r_0 conditional on drawing ω_1. □

B.4 Reputation for Being Not Greedy

Individuals want to signal something about themselves, a particular characteristic. In this case, individuals vary in how much they like money versus their reputation about a specific characteristic (e.g., not being greedy). We use as an inspiration the motivations provided in Bénabou and Tirole (2006) and Fischbacher and Föllmi-Heusi (2013). Individuals with a high reputation want to signal that they are as such. Indexing type by $\theta \in [0, 1]$, the utility is:

$$\phi(r, E(\theta|r); \theta)$$

We assume ϕ is increasing in the first element, i.e., individuals like money; more specifically, the partial is equal to 0 when $1 - \theta = 0$. ϕ is also increasing in the second element, so individuals like others to think they have a high θ; and the partial of ϕ with respect to the second element is 0 when $\theta = 0$. In addition, the cross partial of ϕ with respect to the first element and $1 - \theta$ is positive. This captures the property that individuals face both a higher benefit, and a higher marginal benefit, of the monetary payoff when $1 - \theta$ is larger. Moreover, the cross partial of ϕ with respect to the second element and θ is positive. This captures the property that individuals with higher θs have both a higher benefit, and a higher marginal benefit, of being perceived as having a higher expected θ. Intuitively our assumptions are tantamount to supposing that less "greedy" individuals also care more about being thought of as less greedy. Greedy individuals care less about being thought of as greedy.

Observation 4 *Suppose individuals have Reputation for Being Not Greedy utility. The set of equilibria does not change with F or with the observability of the states. Some individuals will lie downwards when the state is observed or when it is not observed. Last, equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: Importantly, the equilibrium distribution of reports may not be unique. However, the set of equilibria will not shift with F. To see this, first consider an equilibrium that has full support. Conditional on drawing a particular state, individuals will follow a threshold

rule — people with $\theta \geq \bar{\theta}$ will give one report, and everyone else a different report. Second, in any equilibrium with full support there will be a higher average value of θ for r_0 rather than r_1. Third, from our previous observation regarding Generalized Reputation models, the threshold θ must be the same for individuals who draw ω_0 and ω_1 and so some individuals will lie downwards.

Considering the indifferent type; the indifference condition is: $\phi(r_0, E(\theta|\theta > \bar{\theta}); \bar{\theta}) = \phi(r_1, E(\theta|\theta \leq \bar{\theta}); \bar{\theta})$. An equilibrium will exist given the continuity of ϕ and the property that the expectation of θ is continuous in the cutoff $\bar{\theta}$ (although it may be a corner equilibrium without full support on all reports). Observe that whether or not this condition holds does not depend on F. Intuitively, the utility for any given individual does not depend on what they drew but solely on their report and the reports of other individuals. Thus, if the indifference condition, conditional on some equilibrium strategy holds for F it holds for F'.

Similarly, consider an equilibrium without full support. Without loss of generality suppose all individuals take action r_1. Then for this to be an equilibrium no individual will want to deviate: there exists a $\mu \in [0, 1]$ about the average θ of people reporting r_0 such that for all θ $\phi(r_0, \mu; \theta) < \phi(r_1, E(\theta|r_1); \theta)$. Again, as above, this condition does not depend on F.
Moreover, although reports are used to infer something about the individuals, it is not the probability of being a liar (i.e. something that depends on the drawn state). Thus observing the state, as well as the report, will not actually assist with inferring the type of the individual, and again not change the set of possible equilibria and the predictions regarding downwards lying is the same under observability
An equilibrium with full support occurs whenever there exists a partially separating equilibrium, which does not depend on the support of F.□

The uniqueness or not of the equilibrium depends on the shape of H. Moreover, because G does not directly enter into the utility function and there are potentially multiple equilibria, this model makes no clear predictions regarding how $\hat{G}$ should change G.[44]

[44]Even with a unique equilibrium we cannot make a prediction. Shifts in $\hat{G}$ can shift beliefs about $E(\theta|\theta > \bar{\theta})$ (the expected value of θ conditional on reporting r_0) and $E(\theta|\theta \leq \bar{\theta})$ (the expected value of θ conditional on reporting r_1) in arbitrary ways, thus making the shift consistent with affinity or aversion depending on the distribution H.

B.5 RH with Cursedness

In the RH model, we assume that individuals conduct rational inference about their reputations. In this and the next model, we relax the rationality assumption and allow for the decision-maker to be boundedly rational. Analogously, one could suppose we are considering a rational decision-maker who is trying to impress an audience that they know is boundedly rational. In this model, individuals have the same utility function as in RH. Here we specifically consider "cursed" (Eyster and Rabin 2005) individuals. Cursedness is one of the leading ways to model boundedly rational beliefs. In particular, individuals correctly predict the distribution of other players' actions. However, they underestimate the degree to which others' actions are correlated with their private information. In particular, when individuals are fully cursed, the assumption we will maintain, any player supposes that everyone else's individual strategy profile corresponds to the unconditional aggregate strategy profile. In another sense, any given player supposes that all other players' actions are uncorrelated to their private information. Thus, any given player believes that all other players will have the same (mixed) strategy, regardless of their draw; which is correct in equilibrium. In reality, any given player's strategy will depend on their true state, it is just that they believe that no other player's strategy will depend on the draw. The following observation summarizes the predictions of the model.

Observation 5 *Under RH with Cursedness, we observe drawing in and g-invariance. Some individuals will lie downwards in equilibrium when states are unobserved, but never with observability (and so we observe o-shift). Last, an equilibrium with full support can only exist if* $f(\omega_0) > 0.5$.

Proof: Any individual believes that others, regardless of what they drew, have the same best response function (given a distribution F over states). Given an equilibrium with full support and unobservability, some individuals must lie down, for the same reason as the standard RH model. Suppose that the unconditional probability (regardless of the drawn state) of reporting r_1 is α. Then, any given player supposes that all other players, regardless of their draw, have a probability α of reporting r_1. Thus any given individual believes that the fraction of liars reporting r_1 is $\frac{\alpha f(\omega_0)}{\alpha f(\omega_1)+\alpha f(\omega_0)} = f(\omega_0)$ and the fraction of liars reporting r_0 is $\frac{(1-\alpha) f(\omega_1)}{(1-\alpha) f(\omega_1)+(1-\alpha) f(\omega_0)} = f(\omega_1)$. Thus, for the same reasons as the RH model, the reporting

distribution is unique and an equilibrium exists. We also get drawing in for the exact same reason as in the RH model. Again, as in the RH model, the threshold calculation for individuals does not depend on G, but only on F. Thus, the indifferent type will not change with changes in $\hat{G}$.

As in the RH model, if the state as well as the report is known then no inference needs to be done, the probability of being a liar is either 0 or 1, and so behavior should change. Moreover, because the motivation to lie downwards is driven by reputational concerns, when the states are observed, individuals have no motivation to lie down. Moreover, if $f(\omega_0) \leq f(\omega_1)$ there cannot be any equilibrium with full support, again for the same reason as the RH model. □

B.6 RH with Level-k Thinking

The other leading model of boundedly rational beliefs is level-k thinking, as first discussed by Stahl and Wilson (1994). In the RH with Level-k Thinking model, individuals have the same utility function as in RH, but the audience is boundedly rational in the specific sense that they are level-k thinkers. In this class of models, each agent has a level of strategic sophistication k. Level-0 individuals play "randomly" and level-k individuals, for $k > 1$, best respond to level $k-1$ individuals. Thus, individuals have two parameters: θ, which captures the cost of reputation, and their strategic sophistication k.

Individuals' actions will depend on both their k, as well as what they think the distribution of level-0 actions are. Thus, making strong predictions about level-k can be difficult. In order to put structure on the predictions of the model we will assume that all individuals believe that they face a population which puts full support on all reports.

This assumption is a way of making the model consistent with the data. If an individual of level k believed that they faced a population of level $k-1$ which did not place full support on all possible reports, then they would observe data which is inconsistent with their model of the world. In contrast, as our proof will make clear, regardless of their level, subjects can find some belief about H which will rationalize the observed full support distribution over reports.

Believing that level $k-1$ individuals place full support on the reporting distribution implies beliefs about level $k-2$ behavior, and so on, down to level 0. Thus, we need to

make an assumption about level 0 individuals to rationalize all higher level's of behavior. In particular, we assume that level-0 individuals play a mixed strategy of reporting the high and low reports, where the mixing is invariant to the state they drew. [45] Given these assumptions the following observation summarizes the predictions of the model.

Observation 6 *Under RH with Level-k Thinking, we observe drawing in and g-invariance. Some individuals will lie downwards in equilibrium when states are unobserved, but never with observability (and so we observe o-shift). Last, an equilibrium with full support can only exist if* $f(\omega_0) > 0.5$.

Proof: First, consider an individual level $k > 1$ and type θ believing that they are facing a set of level $k-1$ thinkers. Given the assumption of full support, they believe there will be an indifferent type in the level below them; denoted $\bar{\theta}_{k-1}$. Observe that this indifferent type must satisfy the condition $\phi(r_0, \Lambda_{k-1}(r_0); \bar{\theta}_{k-1}) = \phi(r_1, \Lambda_{k-1}(r_1); \bar{\theta}_{k-1})$, where $\Lambda_{k-1}(r_0)$ represents the beliefs of level $k-1$ individuals about the probability of liars at r_0 (and similarly for r_1). As in RH, individuals' optimal reports do not depend on their drawn state (since utility doesn't depend on the state) and so some individuals must be lying down. Thus, using the exact same algebra to solve for the proportion of liars reporting r_0 and r_1 as in the RH model, we can rewrite this condition as $\phi(r_0, 1 - f(\omega_0); \bar{\theta}_{k-1}) = \phi(r_1, f(\omega_0); \bar{\theta}_{k-1})$. Thus, a level-k individual will believe that the set of agents they are best responding to has a threshold that depends only on f, and in a way that is exactly analogous to the RH model.

Given that we suppose that there must also be full support on the reports by level k individuals, this implies that there is again some θ with sophistication level k that will be indifferent between the two actions, so denoting this indifferent type as $\bar{\theta}_k$ then we know $\phi(r_0, \Lambda_k(r_0); \bar{\theta}_k) = \phi(r_1, \Lambda_k(r_1); \bar{\theta}_k)$, or again $\phi(r_0, 1 - f(\omega_0); \bar{\theta}_k) = \phi(r_1, f(\omega_0); \bar{\theta}_k)$. For the same reasons as the RH model, the reporting distribution is unique and exists. We will observe drawing in, just as in the RH model, for the same reason. Moreover, observe that for this type, the threshold calculation for individuals does not depend on G, but only on F. Thus, the indifferent type will not change with changes in $\hat{G}$ for all $k > 1$. The fact that individuals

[45] The other plausible assumption that generates full support for level-0 is that all level-0 individuals report the truth. In this case, all level-1 individuals would report r_1. Thus, the level-2 individuals would not expect to be facing an equilibrium with full support and so have no beliefs about H that could be consistent with the data. Our assumption on level-0 thinkers is equivalent to "as if" the indifferent type for level-0's was equal to exactly the mean of the H distribution.

will lie down in equilibrium without observability of the states, but will not when the states are observable, again follows from the RH reasoning. We also have the same condition as the RH model for a full support outcome

Now, we just have to consider level $k = 1$. The proof here is almost analogous. These individuals believe they are facing a distribution with full support composed of level-0's who randomly choose their report, independent of their drawn state or θ. Thus, a level 1 individual will believe that the set of agents they are best responding to has a threshold that depends only on f, and in a way that is exactly analogous to the RH model. Now, for the individual of level 1, and with full support, observe that there is again some type that will be indifferent between the two actions, so $\phi(r_0, \Lambda_1(r_0); \bar{\theta}_1) = \phi(r_1, \Lambda_1(r_1); \bar{\theta}_1)$, or again $\phi(r_0, 1-f(\omega_0); \bar{\theta}_1) = \phi(r_1, f(\omega_0); \bar{\theta}_1)$. This is analogous to the condition derived in the previous paragraph, and the results are exactly the same. □

B.7 Normalized RH Model

The CLC model (Section 2.4) normalizes the cost of lying by the average cost of lying. One could imagine a similarly "normalized" RH model, which normalizes the reputational cost of a report by the average reputational cost. Utility is:

$$\phi(r, \nu(\Lambda(r_i), \textstyle\sum_j \Lambda(r_j)G(r_j)); \theta)$$

$\Lambda(r)$ has the same interpretation as previously. $\sum_j \Lambda(r_j)G(r_j)$ is the total fraction of liars overall. ν is the normalized cost of reporting r and is increasing in the first argument, and falling in the second. ϕ is strictly increasing in its first argument and falling in the second (and moreover, ϕ does not change with the second argument if $\theta = 0$), and falling in θ. Last, the cross partial of ϕ with respect to ν and θ is negative (strictly so when both arguments are strictly positive).

Observation 7 *Suppose individuals have Normalized RH utility. Some individuals will lie downwards in equilibrium when states are unobserved but never with observability (and so we observe o-shift). An equilibrium with full support can only exist if* $f(\omega_0) > 0.5$.

Proof: As in the RH model, in any equilibrium with full support, there will be a threshold type $\bar{\theta}$, such that individuals with a lower type will report r_0 and individuals with a higher

type will report r_1 (and so we must have lying down when the states are unobserved). Recall that $P(\bar{\theta}) = P(\theta < \bar{\theta})$ is the probability of being less than threshold type $\bar{\theta}$.

The total fraction of liars is $P(\bar{\theta})f(\omega_0)+(1-P(\bar{\theta}))(1-f(\omega_0)) = 2P(\bar{\theta})f(\omega_0)+1-f(\omega_0)-P(\bar{\theta})$ which we denote as Z. The probability of being a liar reporting r_0 is $\frac{f(\omega_1)(1-P(\bar{\theta}))}{f(\omega_1)(1-P(\bar{\theta}))+f(\omega_0)(1-P(\bar{\theta}))} = f(\omega_1)$, and the probability of being a liar reporting r_1 is $\frac{f(\omega_0)(P(\bar{\theta}))}{f(\omega_0)P(\bar{\theta})+f(\omega_1)P(\bar{\theta})} = f(\omega_0)$. The indifference condition is then: $\phi(r_1, \nu(f(\omega_0), Z); \bar{\theta}) = \phi(r_0, \nu(1-f(\omega_0), Z); \bar{\theta})$. Unlike the RH model, the threshold calculation for individuals may now depend on H or G via Z, and in an arbitrary way (and so we may not have uniqueness of the reporting distribution). However, an equilibrium will exist because of continuity of ϕ and the property that Λ and Z are continuous in the threshold θ.

Individuals' behavior will change if the state is observed for analogous reasons as in the RH model, and so with observability there will be no lying down. Observe that, as in the RH model, if $f(\omega_0) \leq f(\omega_1)$ then no equilibrium with full support exists. □

Without additional restrictions on ν no definitive statement regarding whether we observe drawing in, drawing out or f-invariance can be made. Moreover, the indifference condition in an equilibrium with full support is a function of Z, so the indifferent type may change with changes in $\hat{G}$ in either direction, leading to either affinity or aversion or g-invariance.

B.8 Collective Reputation

Individuals may care about not about their individual reputation, but rather the reputation of the entire subject pool as a group. For example, Cohn et al. (2014) study bankers who might be aware of the bad reputation bankers currently have and might try to show to the world that bankers are indeed much more honest than people think.

In this case, deciding whether to lie or not becomes a public good problem. With a sufficiently large group of individuals, everyone has an incentive to maximally lie, because they have a negligible effect on the group's reputation, and so we get the same prediction as the standard model.

B.9 Guilt Aversion

Guilt aversion (Charness and Dufwenberg 2006; Battigalli and Dufwenberg 2007, 2009) has been widely used to understand how individuals report in sender-receiver games. Guilt aversion supposes that individuals feel guilty if they behave in a way that disappoints others, i.e., if their behavior makes the monetary payoff of others fall below their expectation. Our set-up involves only subjects reporting their draw to the experimenter, so one might argue that guilt aversion is not appropriate for this subject-experimenter interaction. We still include it in our list of models since it has been widely applied and we want our study to be able to link to that literature. Moreover, in several experiments surveyed in the meta study (Appendix A), a higher report reduces the payoff of another subject (and not the budget of the experimenter). In those treatments, guilt aversion could well be applied to the subject-subject interaction. Average behavior in these treatments is quite similar to the behavior in subject-experimenter treatments (see Table A.2), so it could well be that similar motives play a role in the subject-experimenter interaction.

In applying guilt aversion to our setting, we thus assume that subjects feel they may disappoint the experimenter if they give too large a report. In particular, a subject feels no guilt if they report less than the average report $\bar{r}$ expected by the experimenter, but if they report more than the expected report they feel guilt proportional to the gap between the expected report and their actual report. Although models of guilt aversion are built on psychological games, in our set-up the model becomes very similar to social comparison models. Utility is:

$$\phi(r, \nu\varsigma(r - \bar{r}))$$

As in the previous model of inequality aversion, ς is the function that maps the difference between any given individual's payoff and the average payoff to a utility cost. If $r \leq \bar{r}$, then $\varsigma(r - \bar{r}) = 0$. If $r > \bar{r}$, then $\varsigma(r - \bar{r})$ is increasing in $r - \bar{r}$. ν is the individual parameter that governs the weight that an individual applies to guilt. We suppose that ϕ is strictly increasing in its first argument and decreasing in its second. Guilt aversion predicts the same as the inequality aversion model, due to the similarity of the functional form and the assumptions. The key difference is in the assumption about utilities when $r \leq \bar{r}$ (which do not affect the predictions). Now, the observed distribution of reports (i.e. the equilibrium

reporting distribution) may not be unique. However because G enters in the utility function directly (because G has a one-to-one mapping with $\bar{r}$) we can still make predictions regarding the effect of $\hat{G}$.

Observation 8 *Suppose individuals have Guilt Aversion utility. The set of equilibria does not change with F, and we observe affinity. The observability of the state does not affect the set of equilibria. Equilibria with full support can exist for any F where $f(\omega_0) \in (0,1)$.*

Proof: See the proof for results on Inequality Aversion. □

B.10 Inequality Aversion + LC

We extend the simple inequality aversion model we develop in Section B.1, so that individuals additionally care about the cost of lying (for an early version of such a model, see Hurkens and Kartik 2009). Formally, utility is

$$\phi(r, \nu\varsigma(r - \bar{r}), c(r, \omega); \theta)$$

where $\bar{r}$ is the mean report. c has the same properties as in the LC model. As before, ν is the individual-specific parameter that governs the weight that an individual applies to social comparisons. We now have a joint distribution over possible values of θ and ν which is non-atomic in the range $[0,1]\times[0,1]$. ς has the same properties as in the inequality aversion model. We suppose that ϕ is strictly increasing in its first argument, strictly decreasing in its second, and decreasing in its third (moreover, as before, the partial is 0 when $\theta = 0$). Moreover, it is decreasing in θ and the cross partial of ϕ with respect to c and θ is negative (strictly so when both arguments are strictly positive).

Observation 9 *Suppose individuals have Inequality Aversion+LC utility. With a unique equilibrium we will observe drawing in, and we will always observe affinity. The observability of the state does not affect the set of equilibria. Last, equilibria with full support can exist for any F where $f(\omega_0) \in (0,1)$.*

Proof: As with the standard inequality aversion model, the observed distribution of reports (i.e. the equilibrium reporting distribution) may not be unique. We can think of the equilibrium as now being characterized by a set of combinations of θs and νs, who conditional on

a drawn state, are indifferent between the two reports. We can think of this set as being a function in the space $\nu \times \theta$; or graphically, a line in two-dimensional Euclidean space. Given a particular utility function, each of these line can be characterized by its θ intercept and its ν intercept, denoted $\theta_T^{\omega_i}$ and $\nu_T^{\omega_i}$. Moreover, since the LC portion of costs never depends on the distribution of responses, the $\theta_T^{\omega_i}$ intercept (i.e. the threshold value of $\theta_T^{\omega_i}$ when $\nu = 0$) must always be the same, and so in fact, we can think of this line as really being characterized by a single intercept: $\nu_T^{\omega_i}$ (for each state ω_1 and ω_2), which is the indifferent type when $\theta = 0$.[46]

Thus, in order to solve for an equilibrium we can consider a function $\zeta(\nu_T^{\omega_1}, \nu_T^{\omega_2})$, which maps from the threshold intercepts that all other individuals are using to the optimal threshold intercepts for any given individual (it maps from R^2 to R^2). Thus, the fixed point(s) of $\zeta(\nu_T^{\omega_i}, \nu_T^{\omega_i})$ characterize our equilibrium (this is true even for multiple equilibria). However, observe that because we are looking at the ν intercepts, this is where the standard lying costs are 0. Thus, the actual drawn state does not enter the utility, and so players must behave the same regardless of which state they drew; so $\nu_T^{\omega_1} = \nu_T^{\omega_2}$. Thus, our problem reduces to a single dimension; and we can consider the function $\zeta(\nu_T)$. An equilibrium will exist because of the continuity of ϕ and ς and the property that $\bar{r}$ is continuous in the threshold types (where the threshold is in ν).

Given this, suppose $f(\omega_1)$ increases. Now consider what happens to $\zeta(\nu_T)$. Observe that because of the shift, for a fixed thresholds that all others are using, there are individuals who previously would have reported r_0 who will now report r_1, namely those with relatively high θs. Thus $\bar{r}$ increases, and the best response of individuals is to increase their reports (regardless of which state they drew). So the thresholds will shift up in both states (conditional on any pair of input thresholds) and at the fixed point we observe drawing in.

Although the equilibrium may not be unique, because G enters in the utility function directly (because G has a one-to-one mapping with $\bar{r}$, but does not affect the lying costs) we can still make predictions regarding the effect of $\hat{G}$. If $\hat{g}(r_1)$ increases, then the beliefs about $\bar{r}$ increases, and so the incentive to say r_1 increases, leading to affinity. As in the standard inequality aversion model, if individuals dislike being too far ahead of others, they may lie down.

[46] To see that this is sufficient to characterize the equilibrium, observe that there is a one to one mapping between $\bar{r}$ and the indifferent type when $\theta = 0$.

With sufficiently low inequality costs and sufficiently high lying costs, individuals will not lie, so there exist equilibria with full support for any F where $f(\omega_0) \in (0,1)$. In addition, whether there is lying downwards will depend on the function ς just as in the inequality aversion model. Moreover, as in the LC model and the inequality aversion model, the distribution of reports does not depend on observability. □

B.11 Keeping Up with the Joneses

One major motivation in social comparison models is a desire not to fall too far behind one's comparison group. We capture this intuition in this model by a concern for how many people have "moved ahead of you" or "moved behind you" in the distribution of reports compared to the distribution of states. The reference point for the comparison of monetary payoffs is thus the hypothetical distribution if everyone reported truthfully. We call it the "Keeping Up with the Joneses" model.

Utility is

$$\phi(r, SC; \theta)$$

where SC is the utility due to social comparisons. θ parameterizes how much individuals care about social comparisons, relative to their monetary payoff. ϕ is strictly increasing in the first argument, decreasing in the second (as before, the partial is 0 when $\theta = 0$) and decreasing in θ. Moreover, ϕ has a negative cross partial derivative in SC and θ (strictly negative when both arguments are strictly positive).We now turn to specifying the SC term of the utility function.

We denote $\varsigma(\omega, r)$ as the fraction of people who drew ω and are reporting r (in equilibrium). Because we have only two possible states and reports, there are three possibilities that may happen when an individual compares their place in the drawn state distribution to the their place in the distribution of reports: i) their relative comparison to someone else doesn't change from the state distribution to the report distribution ii) they improve in comparison to someone else when moving from the state to the report distribution (for example, they move from behind to tied) or iii) they become worse off in comparison to someone else when moving from the state to the report distribution.

The utility weight that a person places on someone else moving from tied with them

to different is γ_1. The utility weight that individuals places on someone else moving from different from them to tied is γ_2. Moreover, if the change represents an improvement to that individual (relative to the other person) they experience shame (or guilt), which has weight λ_s. In contrast, if that individual falls behind someone else, they experience envy with a weight λ_e. If someone draws ω_0 and reports r_0, they cannot experience any shame, because they haven't moved ahead of anyone. But they can experience envy at those people drew ω_0 and now report r_1. Their utility is $\phi(r_0, \gamma_1\lambda_e\varsigma(\omega_0, r_1); \theta)$ where $\varsigma(\omega_0, r_1)$ represents the proportion of people drawing ω_0 and saying r_1. Similarly, if someone draws ω_1 and reports r_1, they cannot experience any shame, because they haven't moved ahead of anyone. But they can experience envy, at those people drew ω_0 and now report r_1. Their utility is $\phi(r_1, \gamma_2\lambda_e\varsigma(\omega_0, r_1); \theta)$.

In order to understand the model a bit better, consider the decision of whether to lie or not by an individual drawing either state. Observe that no-one who draws ω_1 will report r_0. This is because it implies that the individual would fall behind others in the report distribution relative to the state distribution, incurring a positive cost. Moreover, it would give a lower monetary payoff. If someone draws ω_0 and reports r_1, they cannot experience any envy, because they haven't fallen behind anyone. But they can experience shame, at those people drew ω_0 and now report r_0, and at those people who drew ω_1 and report r_1 (they would also feel shame towards those who drew ω_1 and report r_0, if there were any). Their utility is $\phi(r_1, \gamma_1\lambda_s\varsigma(\omega_0, r_0) + \gamma_2\lambda_s\varsigma(\omega_1, r_1); \theta)$.

Observation 10 *Suppose individuals have Keeping up with the Joneses utility. We observe affinity. The observability of the state does not affect the set of equilibria. Individuals never lie downwards. Equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: We discussed previously that individuals will not lie down. Given an equilibrium with full support the indifferent type who drew ω_0 must satisfy

$$\phi(r_1, \gamma_1\lambda_s\varsigma(\omega_0, r_0) + \gamma_2\lambda_s\varsigma(\omega_1, r_1); \theta) = \phi(r_0, \gamma_1\lambda_e\varsigma(\omega_0, r_1); \theta)$$

As with the previously discussed social comparison models the equilibrium reporting distribution may not be unique. An equilibrium will exist because of the continuity of all of the functions and the property that G is continuous in the threshold type.

Even though the equilibrium may not be unique, because G enters in the the utility function directly (via the number of individuals reporting r_1 and r_0) we can still make predictions regarding the effect of $\hat{G}$. Suppose that, fixing F, $\hat{g}(r_1)$ increases. Then $\varsigma(\omega_0, r_1)$ must have increased, and $\varsigma(\omega_0, r_0)$ must have decreased. So the utility from reporting r_0 has fallen, and the utility from reporting r_1 has increased, so the indifferent type now wants to lie to r_1.

Since no part of the utility function depends on observability, making the state observable does not affect the set of possible equilibria (or our prediction about downwards lying). Full-support equilibria can exist for any F where $f(\omega_0) \in (0, 1)$ as long as an individual experiences enough shame relative to monetary benefits. □

Even with a unique equilibrium we may observe either drawing in or drawing out or f-invariance. We know that indifferent type who drew ω_0 must satisfy

$$\phi(r_1, \gamma_1 \lambda_s \varsigma(\omega_0, r_0) + \gamma_2 \lambda_s \varsigma(\omega_1, r_1); \theta) = \phi(r_0, \gamma_1 \lambda_e \varsigma(\omega_0, r_1); \theta)$$

Now suppose we increase $f(\omega_1)$. Observe that fixing the indifferent type, this implies that $\varsigma(\omega_1, r_1)$ has increased by the change in $f(\omega_1)$ and that both $\varsigma(\omega_0, r_0)$ and $\varsigma(\omega_0, r_1)$ have fallen by less than that amount. Thus, the utility of reporting r_0 has increased. However, the utility of reporting r_1 may decrease, or increase (by even more than the utility or reporting r_0) depending on the relative values of the γs and the λs.

B.12 Conformity in Actions

In the Conformity in Lying Cost (CLC) model (Section 2.4), we formalize the effect of social norms by assuming that individuals want to conform to others in the amount of lying they do. Here, we discuss a model of "Conformity in Actions" in which individuals' utilities depend on their report, and also how closely their report aligns with everyone else's report. The utility function is thus only a function of any individual's report and the distribution of reports:

$$\phi(r, G; \nu).$$

We suppose that utility improves as any given individual conforms more to the modal report. Thus,fixing a report, the utility from giving that report (relative to the other report) is increasingas G puts more weight on that report.Moreover, individuals vary in their cost of

deviating from the modal report, where the cost is increasing in ν (which is distributed on [0,1]).

Observation 11 *Suppose individuals have Conformity in Actions utility. The set of equilibria does not change with F and we observe affinity. The observability of the state does not affect the set of equilibria. Last, equilibria with full support can exist for any F where $f(\omega_0) \in (0,1)$.*

Proof: First, observe that the set of equilibria will not change with F. This is because any given individual's utility does not depend on their drawn state, but rather on their own report and everyone else's report. Thus, any individual's utility from any given report only depends on the distribution of others' reports. The observed distribution of reports (i.e. the equilibrium reporting distribution) may not be unique, although given our continuity assumptions an equilibrium will exist. However, because G enters in the the utility function directly (because G has a one-to-one mapping with the modal report) we can still make predictions regarding the effect of $\hat{G}$. Thus, by assumption, as $\hat{g}(1)$ increases, the incentive to say r_1 is increasing. Since no part of the utility function depends on observability, making the state observable does not change behavior. The model can generate lying downwards if individuals' desire to conform to others is strong enough. The model can generate full support for any F where $f(\omega_0) \in (0,1)$, so long as r_0 is the modal report, but some individuals deviate to receive a higher monetary payoff. □

B.13 Censored Conformity in Lying Costs

This section presents a variation of the CLC model. One could imagine that an individual does not normalize their lying cost by the average lying cost in society (as in CLC), but only by the lying costs incurred by individuals who "could have" lied profitably, i.e. those who did not receive the maximal draw.

In this model, as in CLC, individuals will not want to lie downwards. In order to simplify notation in terms of the utility function, we will directly assume no one lies downwards. The utility function is then:

$$\phi(r, \eta(c(r,\omega), \tfrac{g(r_1)-f(\omega_1)}{f(\omega_0)}); \theta)$$

where η is the normalized cost function. Observe that $\frac{g(r_1)-f(\omega_1)}{f(\omega_0)}$ represents the share of liars among those who drew ω_0 (and so is distinct from the test statistic we discuss in the body of the paper). η is falling in the first argument and increasing in the second. We suppose that $\eta(0, \frac{g(r_1)-f(\omega_1)}{f(\omega_0)}) = 0$. ϕ is strictly increasing in the first argument, falling in the second (where, specifically, the partial is 0 when $\theta = 0$), and falling in θ. Last, the cross partial of ϕ with respect to η and θ is negative (strictly so when both arguments are strictly positive).

Observation 12 *Suppose individuals have Censored Conformity in Lying Costs utility. We observe affinity. The observability of the state does not affect the set of equilibria. Individuals never lie downwards. Equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: Consider an equilibrium with full support. There must be an indifferent type who drew ω_0: $\phi(r_1, \eta(c(r_1, \omega_0), \frac{g(r_1)-f(\omega_1)}{f(\omega_0)}); \bar{\theta}) = \phi(r_0, \eta(c(r_0, \omega_0), \frac{g(r_1)-f(\omega_1)}{f(\omega_0)}); \bar{\theta})$. Recall that given a $\bar{\theta}$ we say $P(\bar{\theta}) = P(\theta < \bar{\theta})$ is the probability of θ being less than threshold type $\bar{\theta}$. We can then rewrite the indifference condition as $\phi(r_1, \eta(c(r_1, \omega_0), P(\bar{\theta})f(\omega_0)) : \bar{\theta}) = \phi(r_0, \eta(c(r_0, \omega_0), P(\bar{\theta})f(\omega_0)) : \bar{\theta})$.

As in the standard CLC model, the equilibrium reporting distribution may not be unique. An equilibrium will exist given the continuity of ϕ and η and the property that $\bar{c}$ is continuous in the cutoff $\bar{\theta}$ (although it may be a corner equilibrium without full support on all reports).

Although we do not necessarily have a unique equilibrium, since G directly enters in the utility function we can still make predictions regarding the effect of $\hat{G}$. Thus, suppose we fix F, and the indifferent type. Now allow $\hat{g}(r_1)$ to increase. Then there must be more liars who drew ω_0 and so the second argument of the utility function must increase. Thus, the cost of lying goes down, and the previously indifferent type must strictly prefer to lie. Since no part of the utility function depends on observability, making the state observable does not change behavior. Individuals will not lie down for the same reason as in the CLC model. Full support equilibrium can exists when for any F where $f(\omega_0) \in (0, 1)$ for the same reason as the CLC model. □

We may observe either drawing in or drawing out or f-invariance. To see this, consider what happens when $f(\omega_0)$ falls. As with the CLC model we consider the function $\zeta(\bar{\theta})$, which maps from Θ to Θ and has the interpretation: given a threshold $\bar{\theta}$ that all other individuals are using, $\zeta(\bar{\theta})$ gives the optimal threshold. Then, the fixed point(s) of $\zeta(\bar{\theta})$ characterizes our

equilibrium (this is true even for multiple equilibria). Observe that fixing everyone else's $\bar{\theta}$, as $f(\omega_0)$ falls the second argument of η may rise or fall. Thus lying may become more or less attractive and so we may have drawing in or out.

B.14 Kőszegi-Rabin + LC

We can also combine the intuition of the Kőszegi-Rabin model with the lying cost model. We suppose that individuals face lying costs and experience gain-loss utility both over monetary outcomes, and over the lying costs (possibly to different degrees). As before we will denote the cost of reporting r if ω is the state as $c(r, \omega)$ which has the same properties as described under LC. The utility of reporting r if ω is the state is then:

$$\phi(r, \omega, a; \theta, \rho, \lambda_u, \lambda_c) =$$
$$u(r) - \theta c(r, \omega) + \rho[\textstyle\sum_k \lambda_u^{\mathbb{I}} |u(r) - u(a(\omega_k))| f(\omega_k) + \theta \sum_k \lambda_c^{\mathbb{I}} |c(a(\omega_k), \omega_k) - c(a(\omega), \omega)| f(\omega_k)]$$

θ parameterizes the cost of lying. ρ is a scalar representing the weight on gain-loss utility, and λ_u and λ_c represent the separate gain-loss parameters for money and lying costs. $\lambda_i^{\mathbb{I}}$ is an indicator function that takes on a value of 1 if the argument inside the attached absolute value is positive, and $-\lambda$ otherwise. An equilibrium a is a mapping such that if a maps $\hat{\omega}$ to $\hat{r}$, then the argmax of $\phi(r, \hat{\omega}, a; \theta, \rho, \lambda_u, \lambda_c)$ is $\hat{r}$. As Kőszegi and Rabin point out, there may be multiple "personal equilibria" mappings a. However, there will generically be a unique preferred personal equilibrium, i.e. an equilibrium mapping a that gives the highest utility, among all possible equilibrium as for any given value of λ_u and λ_c. We will suppose, in line with Kőszegi and Rabin (2006, 2007), that individuals choose the preferred personal equilibrium. Then the aggregate distribution of reports is simply the set of reports generated by the distribution of states and as that each individual uses.[47]

Observation 13 *Suppose individuals have Kőszegi-Rabin+LC utility. We observe g-invariance and o-invariance. Individuals never lie downwards. Equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: As in the LC model, no individual will lie downwards (as they will face larger lying costs, and a lower monetary payoff, compared to telling the truth), this model can generate

[47] As Kőszegi and Rabin (2007) point out, such an an a will exist so long as we allow the action space to be convex, i.e. allow randomization.

full support for any F where $f(\omega_0) \in (0,1)$ (so long as lying costs are strong enough), and observability will not affect reports.

However, any individual's strategy, fixing F, will not depend on the distribution of reports in the population: the set of equilibrium mappings is constant in G. Intuitively, it is the case that an individual's expectations of their draw, and their report, depends only on F, not on G. Moreover, any individual's expectations only depend on their draw, and the equilibrium mapping a, but neither of these depends on G. Thus a itself cannot depend on G and thus not on $\hat{G}$. $\square$

This model can exhibit drawing in, drawing out or f-invariance. For example, suppose that an individual exhibits only gain-loss utility in the monetary dimension, but not in the lying cost dimension. Then an increase in $f(\omega_1)$ will increase expectations of monetary payoff, and so, conditional on drawing ω_0, an individual will be more likely to report r_1. In contrast, if an individual exhibits gain-loss utility only in the cost dimension, but not in the monetary dimension, the opposite intuition will be true.

We also conducted an experiment to test some specific predictions of this model, following closely the design of Abeler et al. (2011). Subjects' reports were only paid out with 50 percent probability, and with the other 50 percent subjects received a fixed payment which differed by treatment. This payment lottery was only resolved after subjects made their report. The fixed payment thus enters expectations and the Kőszegi-Rabin + LC model predict that subjects will lie more if the fixed payment is higher. We find, however, no significant difference between treatments (155 subjects, p=0.676, OLS with robust SE).

B.15 General RH+LC

Here we consider a general form of the Reputation for Honesty plus Lying Cost (RH+LC) model. In contrast to the Separable RH+LC model (Section 2.5), this general version allows for non-separability between the components of the utility function. Utility is:

$$\phi(r, c(r,\omega), \Lambda(r); \theta_1, \theta_2)$$

The assumptions on the model are less restrictive than for the Separable RH+LC model. ϕ is strictly increasing in the first argument, but decreasing in the second and third; so that the

individual likes more money, but dislikes lying and being perceived as a liar. The θs indicate the relative weight of the second and third arguments. ϕ is decreasing in both θs. the cross partials of ϕ with respect to c and θ_1 and with respect to Λ and θ_2 are both negative (and strictly so when both respective arguments are strictly positive). As before, these assumptions capture the property that higher θs indicate both a higher cost and marginal cost of lying. If $\theta_1 = 0$, the partial of ϕ with respect to the second argument is 0; and if $\theta_2 = 0$, the partial of ϕ with respect to the third argument is 0. We suppose a joint distribution H over $\theta_1 \times \theta_2$ which is non-atomic.

Since the model imposes very little restrictions and allows for two-dimensional heterogeneity, it does not make specific predictions for our F and $\hat{G}$ treatments. As with the Separable RH+LC model multiple equilibria may occur. Even assuming an unique equilibrium, the model may exhibit drawing in, drawing out or f-invariance, depending on how the set of indifferent types change. Moreover, given that equilibria are not necessarily unique, we cannot predict the effect of the $\hat{G}$ treatments, and even if the equilibrium were unique, we could observe either aversion, affinity or g-invariance. The following observation summarizes our predictions.

Observation 14 *Suppose individuals have General RH+LC utility. We observe o-shift. Individuals may lie downwards when the state is unobserved, but will not when states are observed. Equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: Given sufficiently high lying costs, the model can always generate a distribution of reports with full support (as with the LC model) for any F where $f(\omega_0) \in (0, 1)$.
In order to see how to construct an equilibrium with full support observe that there must be an indifferent type who drew ω_0: $\phi(r_1, c(r_1, \omega_0), \Lambda(r_1); \theta_1, \theta_2) = \phi(r_0, c(r_0, \omega_0), \Lambda(r_0); \theta_1, \theta_2)$; and an indifferent type who drew ω_1: $\phi(r_1, c(r_1, \omega_1), \Lambda(r_1); \theta_1, \theta_2) = \phi(r_0, c(r_0, \omega_1), \Lambda(r_0); \theta, \theta_2)$

Observe that there is a threshold function for each state $\tau_{\omega_i}(\theta_1, \theta_2)$.[48] If τ is less than or equal to some constant, the individual will report their state, otherwise they will lie. So, we can think of the equilibrium as being characterized by a set of combinations of θ_1s and θ_2s. We can think of this set as being a function in the space $\theta_1 \times \theta_2$; or graphically, a line in two-dimensional Euclidean space (see Figure B.1).

[48] These will be functions because of our assumptions on the utility function (i.e. the cross partials of ϕ) and will be continuous because of the continuity of ϕ.

In any equilibrium, r_1 has to have more liars. Suppose no one lies down. Then clearly r_1 has more liars. Now suppose people do lie down, and r_1 has fewer liars than r_0. In this case, consider the individuals whose state is ω_1. They would obtain a better reputation, lower lying costs and a higher monetary payoff, by simply reporting r_1. So r_1 must have more liars.

Thus, fixing a θ_1, as θ_2 increases, an individual's value of reporting r_0 increases. Fixing θ_2, as θ_1 increases, individuals' value of reporting what they drew increases. Moreover, if an individual draws ω_0 and reports r_1 then an individual with the same preference parameters, but with a draw ω_1, must also report r_1. This is because saying r_1 gives the same reputational value and the same monetary payoff to both individuals but the individual who drew ω_0 pays a lying cost. For the same reason, if an individual draws ω_1 and reports r_0 then an individual with the same preference parameters, but with a draw ω_0 must also report r_0.

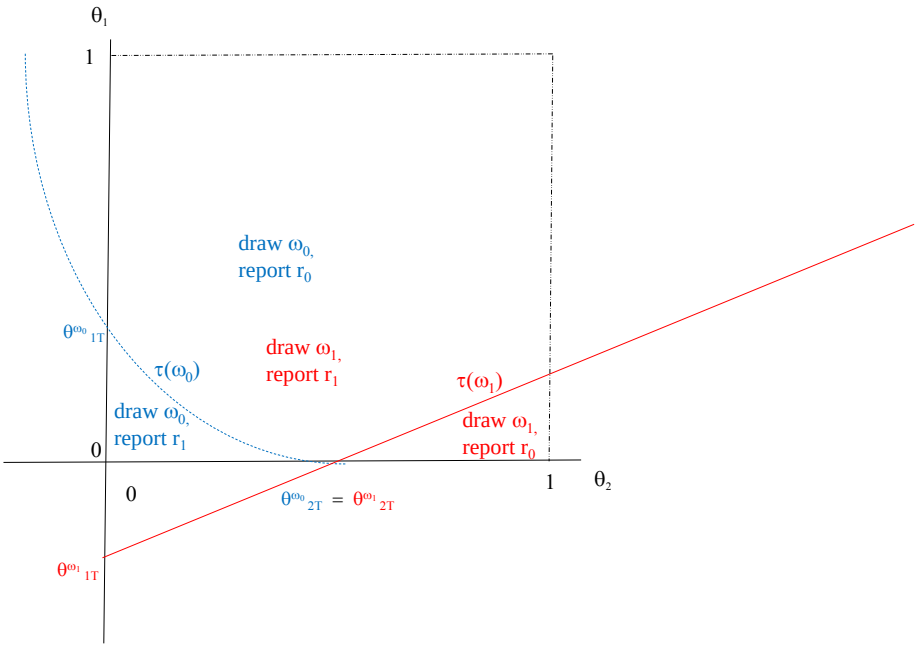


Figure B.1: Thresholds for General RH+LC Model

The equilibrium reporting distribution is not necessarily unique. In order to solve for an equilibrium we can consider a function $\zeta(\Lambda(r_0), \Lambda(r_1))$ which maps from the probability of a liar at any given report to a new set of probabilities of lying at any given report. The way to interpret this is: given some inputs, which represent the fraction of liars at r_0 and r_1, we can construct the best response function for any given individual (given θ_1 and θ_2).

We can use these best responses, along with H and F, to construct a new distribution G and new Λs, which are the output of the function. Thus, the fixed point(s) of $\zeta(\Lambda(r_0), \Lambda(r_1))$ characterize our equilibrium (this is true even for multiple equilibria). Given our continuity assumptions, the threshold functions will be continuous in the Λs, and Λs will be continuous in the threshold functions, so an equilibrium will exist.

Observability will matter as long as some individuals care about the reputation costs. In particular, reputational concerns will no longer drive individuals to report anything other than either the truth or the high state. We will observe no lying downwards at all under observability of the state by the experimenter since doing so would incur an LC cost and a reputational cost. Without observability of the state, we may either observe lying downwards or not. It depends on the threshold functions. If, for example, the horizontal (θ_2) intercept of the threshold function is greater than or equal to 1, then individuals will not lie downwards in equilibrium. Otherwise they will.□

Because, as the proof points out, our equilibrium is found by considering the fixed point of a mapping from a vector of size two to itself, the fixed point problem is in two dimensions. Thus, even given a unique equilibrium we cannot solve for unambiguous comparative statics with shifts in F.[49]

In addition, given that equilibria are not necessarily unique, we cannot predict the effect of the $\hat{G}$ treatments. Moreover, even if the equilibrium of the reporting distribution is unique, we could observe either aversion, affinity or g-invariance. Suppose $\hat{g}(r_1)$ increases. There needs to be more people above the threshold for at least one type. From Figure B.1 we can see that two different shifts in the distribution of θ's could both lead to an increase in $\hat{g}(r_1)$ but one could lead to more liars at r_1 and so drive aversion, but the other could lead to fewer liars at r_1 and so lead to affinity. To see the first shift, suppose that we have a shift in H that shifts mass from above $\tau(\omega_0)$ to below it. This shift doesn't change the reporting of individuals who drew ω_1, but leads to a higher mass of individuals drawing ω_0 to report r_1. This increases $g(r_1)$ but also increases the number of liars at r_1 and will so drive aversion. In contrast, now suppose that we have a shift in H that shifts mass from below $\tau(\omega_1)$ to above it. This shift doesn't change the reporting of individuals who drew ω_0, but leads to a higher

[49] Unlike in the Separable RH+LC model, looking at the intercepts of the threshold functions is not sufficient to pin down the equilibria, as these intercepts only pin down the difference between $\Lambda(r_1)$ and $\Lambda(r_0)$; not their actual values.

mass of individuals drawing ω_1 to report r_1. This increases $g(r_1)$ but also reduces the number of liars at $r_1$and will so drive affinity.

B.16 LC-Reputation

Rather than caring about the reputation of having reported truthfully conditional on their report (the RH model), individuals may instead want to cultivate a reputation as a person who has high lying costs. Such a model is similar to the one discussed in Frankel and Kartik (2016). It is also similar in spirit, although in an entirely different domain, to the models of fairness by Levine (1998), Bénabou and Tirole (2006), Ellingsen and Johannesson (2008), Andreoni and Bernheim (2009), Tadelis (2011), and Grossman (2015). In those models individuals like to be perceived as fair as well as actually having preferences for fairness.

Utility is

$$\phi(r, c(r, \omega), E[\theta_1|r]; \theta_1, \theta_2)$$

c and θ_1 have the same interpretation as in the LC model, and the assumptions regarding them are the same. θ_2 represents the weight that any given individual places on the utility from reputation. Individuals' utility is increasing in $E[\theta_1|r]$ and in θ_2. If$\theta_2 = 0$, the partial of ϕ with respect to the third argument is 0. Moreover the cross partial of ϕ with respect to $E[\theta_1|r]$ and θ_2 is positive (strictly so when both arguments are strictly positive). The interpretation is that individuals have a positive utility from others believing that they have high lying costs. Like the General RH+LC model, the LC-Reputation model fails to make specific predictions for our F and $\hat{G}$ treatments.

Observation 15 *Suppose individuals have LC-Reputation utility. We observe o-shift. Equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: As with LC models, full support equilibria can exist with sufficiently high LC costs for any F where $f(\omega_0) \in (0, 1)$. So long as individuals with either state give both reports, there must be an indifferent type who drew ω_0:

$$\phi(r_1, c(r_1, \omega_0), E[\theta_1|r_1]; \theta_1, \theta_2) = \phi(r_0, c(r_0, \omega_0), E[\theta_1|r_0]; \theta_1, \theta_2)$$

and an indifferent type who drew ω_1:

$$\phi(r_1, c(r_1, \omega_1), E[\theta_1|r_1]; \theta_1, \theta_2) = \phi(r_0, c(r_0, \omega_1), E[\theta_1|r_0]; \theta_1, \theta_2)$$

Observe that there is a threshold function for each state $\tau_{\omega_i}(\theta_1, \theta_2)$, which denotes the set of types that are indifferent between the two reports. If τ is less than or equal to some constant, the individual will report their state, otherwise they will lie.

Also in any equilibrium, reporting r_1 has to provide individuals with a lower inferred (estimated) value of θ_1, compared to reporting r_0. Suppose no one lies down. Then clearly r_1 has a lower estimate of θ_1. Now suppose people do lie down, and r_1 has a higher estimate of θ_1 than r_0. For this case, consider the individuals whose state is ω_1. They would obtain a better reputation, lower lying costs and a higher monetary payoff, by simply reporting r_1. So r_1 must have a lower expected θ_1.

Thus, fixing a θ_1, as θ_2 increases, the individual's value of reporting r_0 increases. Fixing θ_2, as θ_1 increases, individuals' value of reporting what they drew increases. Moreover, if an individual draws ω_0 and reports r_1 then an individual with the same preference parameters, with a draw ω_1 must also report r_1 because saying r_1 gives the same reputation value, the same monetary payoff but the latter pays a lower lying cost. For the same reason, if an individual draws ω_1 and reports r_0 then an individual with the same preference parameters, with a draw ω_0 must also report r_0.

Recall that we can think of the equilibrium as being characterized by a set of combinations of θ_1s and θ_2s. Or equivalently, there is a threshold function for each state $\tau_{\omega_i}(\theta_1, \theta_2)$.[50] If τ is less than or equal to some constant, the individual will report their state, otherwise they will lie. Thus the threshold diagram looks qualitatively similar to Figure B.1, described in Appendix B.15.

The equilibrium reporting distribution may not necessarily be unique. In order to solve for an equilibrium we can consider a function $\zeta(E[\theta_1|r_0], E[\theta_1|r_1])$, which maps from the expected values of θ_1, conditional on reports, to themselves. The way to interpret it is: given some inputs, which represent the expected values of θ_1 at r_0 and r_1, we can construct the

[50] These will be functions because of our assumptions on the utility function (i.e. the cross partials of ϕ) and will be continuous because of the continuity of ϕ.

best response function for any given individual (given their θ_1 and θ_2). We can use these best responses, along with H and F, to construct a new distribution G and new conditional expected values of θ_1; which are the output of the function. Thus, the fixed point(s) of $\zeta(E[\theta_1|r_0], E[\theta_1|r_1])$ characterizes our equilibrium (this is true even for multiple equilibria). Given our continuity assumptions, the threshold functions will be continuous in the conditional expectations of θ_1, and the conditional expectations will be continuous in the threshold functions, so an equilibrium will exist.

Individuals' behavior should change if the state is observed. But this is for a very different reason compared to the RH+LC model. In that model, behavior changes because the probability of being a liar would either be 0 or 1. In the LC Reputation model observing both the state and the report can give a more precise estimate of the cost of lying θ_1, as it can be estimated using both ω and r, rather than just r.

Given the similarity to the RH+LC model, it is clear why lying downwards may occur when states are not observed (and so solely private information). However, lying downwards may still occur in equilibrium when states are observed. This is because the inference is done not on the probability of being a liar, as in RH type models, but about θ_1. It is possible to have a countersignalling equilibrium where the highest and lowest $\theta_1$types pool on truth-telling and middle θ_1 types lie down. □

As with the General RH+LC model, because the fixed point problem is in two dimensions, even given a unique equilibrium we cannot solve for unambiguous comparative statics with shifts in F.[51] Moreover, because the threshold characteristics look qualitatively similar to Figure B.1 we can again see how a shift in $\hat{g}(r_1)$ can cause either affinity, aversion or g-invariance even when the reporting distribution in equilibrium is unique. However, again in this case uniqueness is not necessarily guaranteed depending on the exact functional form of utility.

[51] This is because, as with the general RH+LC model, and unlike the separable version of that model, looking at the intercepts of the threshold functions is not sufficient to pin down the equilibria. This is because these intercepts only pin down the difference between the conditional expectations of θ_1; not their actual values.

B.17 Audit Model

The Audit model builds on the intuition that individuals fear to be "found out" as liars at different rates depending on their report. Individuals who give a report where there are many liars are more likely to be found out as a liar. This may be a concern about an actual audit or, our preferred interpretation, a more metaphorical audit: subjects care about their reputation of being honest – but only if they lied up. If they were honest or lied down, they have a "clean conscience", even though they won't be able to prove their honesty by showing their true state.[52] Another way of interpreting this model is to see it as a combination of a reputation for honesty (like the RH model) with an infinite cost of lying down. Townsend (1979) discusses wanting to avoid detection, which could be motivated by not wanting to be in a category which is likely populated by many liars. Kajackaite and Gneezy (2015) also discuss the possibility of concerns about audits driving results.

Here, individuals are investigated with a probability that is increasing in the proportion of liars that report the same r that they do, i.e., $\Lambda(r)$. If an individual is investigated, and discovered to have been lying upwards they face a utility cost (in our simple binary model it does not matter if this is a fixed cost or proportional to the size of the lie). Individuals face no cost if they are discovered to have lied downwards or have been honest. Individuals' utility function is then

$$\phi(r, \mathbb{I}_{r>\omega}\Lambda(r); \theta)$$

where $\mathbb{I}_{r>\omega}$ is an indicator function which equals 1 if the individual lied upwards, and 0 if the individual did not lie upwards. $\Lambda(r)$ has the same interpretation as previously. We assume that ϕ is strictly increasing in the first argument, and falling in the second argument; and more specifically, the partial of ϕ with respect to the second argument is 0 if $\theta = 0$. And, as before, the cross partial of the second argument and θ is negative (strictly so when both arguments are strictly positive).

Observation 16 *Suppose individuals have Audit utility. We observe drawing in as well as aversion and o-shift. Individuals never lie downwards. Equilibria with full support can exist*

[52] If the audit is an actual concern about the researcher, then there are ways to alleviate such concerns, e.g., by conducting the experiment over the phone or asking subjects to report their prediction about a random draw instead of the draw itself. Our meta study finds only small differences in behavior when the experiment is conducted remotely and mixed results on the effect of reporting an internal state of mind (see Table A.2).

for any F where $f(\omega_0) \in (0,1)$.

Proof: So long as an individual reports at most their actual state, then their utility is a function only of their report. Thus, no individual will lie down, since by telling the truth, they would receive a higher r and no higher costs. This is true with or without observable states.

Next, consider an equilibrium with full support. The individual who drew ω_0 and has a type $\bar{\theta}$ such that he is indifferent between reporting r_0 or r_1 will satisfy: $\phi(r_0, 0; \bar{\theta}) = \phi(r_1, \Lambda(1); \bar{\theta})$.

Thus, the reporting equilibrium is unique because there will be a single type that is indifferent (due to the assumption on the cross partial of Λ and θ). Observe that when $f(\omega_1)$ increases, fixing the indifferent type, the proportion of liars at r_1 must fall. Thus, the type that was previously indifferent will strictly prefer to say r_1. This implies drawing in.

To see that we observe aversion, notice that fixing F, increasing $\hat{g}(r_1)$ implies that there are more liars at r_1. Thus the costs of lying to r_1 (from ω_0) rise, and so fewer individuals are willing to lie.

Last, individuals' behavior will change if the state is observed for the same reasons that it changes in the RH model. Moreover, we can have full support equilibrium for any F where $f(\omega_0) \in (0,1)$ so long as some individuals care enough about being thought a liar.□

B.18 CLC+LC Model

We have considered the combination of RH and LC and we can also consider the combination of CLC and LC. In the CLC+LC model, individuals experience both lying costs and CLC costs. In fact, careful inspection of the assumptions and functional form shows that such a model has the same general functional form and assumptions as the CLC model we considered in body of the paper, and so equivalent predictions. This is due to the fact that CLC itself already includes some version of lying cost.

C Formal Results for Models in the Main Text

C.1 Lying costs (LC)

Observation 17 *Suppose individuals have LC utility. We observe f-invariance, g-invariance and o-invariance. Individuals never lie downwards. Equilibria with full support can exist for any F where $f(\omega_0) \in (0,1)$.*

Proof: First, we show that individuals never lie down. This follows directly from the assumption that $\phi(r, c(r,\omega);\theta)$ is increasing in its first argument and that c has a minimum at $r = \omega$. Thus, fixing a state, individuals always want to report up to the state that they drew (but some who draw ω_0 may report r_1).

We now prove f-invariance. First note that, so long as some individuals experience sufficiently high lying costs, they will not lie, and so the equilibrium (in this case the "equilibrium" is trivial, as any individual's report is independent of others' reports) will feature full support for any F where $f(\omega_0) \in (0,1)$. In a full support equilibrium, we know that those who draw ω_1 will report r_1. Of those who draw ω_0 and report r_1, there will be a cutoff type $\bar{\theta}$, such that $\phi(r_0, c(r_0,\omega_0);\bar{\theta}) = \phi(r_1, c(r_1,\omega_0);\bar{\theta})$. Observe that this cutoff type must be unique by the assumption regarding the cross partial of c and θ. Individuals with $\theta < \bar{\theta}$ who draw ω_0 will report r_1. Individuals with $\theta > \bar{\theta}$ who draw ω_0 will report r_0. Thus, $1 - \frac{g(r_0)}{f(\omega_0)} = 1 - \frac{\text{Prob}(\theta>\bar{\theta})f(\omega_0)}{f(\omega_0)} = 1 - \text{Prob}(\theta > \bar{\theta}) = \text{Prob}(\theta < \bar{\theta})$. This does not depend on F.

Third, the fact that an individual's utility does not depend on G in any way allows us to immediately observe that it exhibits g-invariance. Last, as discussed in the body, our interpretation of LC lying costs are that they are internal costs that are paid and so they do not depend on the inference others are making about any given person. Thus, individuals should not care whether their state was observed. Thus we observe o-invariance and the prediction regarding lying downwards is the same for observable or unobservable states.

The generated reporting distribution is unique. Clearly everyone who draws the high state has a unique best response. Similarly, as described above anyone drawing ω_0 has a unique best response other than type $\bar{\theta}$ who have zero measure. □

We also report the results of an experiment where there are many possible states and reports, rather than just two. We can generalize the intuition from the previous observation to

this situation. We now suppose there are a number of states $\omega_0, \omega_1, ...\omega_N$ and a corresponding number of actions $r_0, r_1, ...r_N$. We can show that if we change the distribution of F, but only for the highest M states, then the distribution of reports will not change for the lowest $N+1-M$ states.. Essentially changes in F for the highest states do not cause changes in G for lower states/reports.

Observation 18 *Under LC, consider two distributions F and F' such that $f(\hat{\omega}) = f'(\hat{\omega})$ for all $\hat{\omega} \leq \omega^*$. Then individuals never lie downwards and for all $\hat{r} \leq r^* = \omega^*$: $g(\hat{r}) = g'(\hat{r})$ for all $(\hat{r}) \leq r^*$. Moreover, we observe g-invariance, o-invariance and we can have an equilibrium with full support when $f(\omega_1) \geq f(\omega_0)$.*

Proof: No individual lies downwards for the same reasons as in the previous observation. Second, observe that the best response of any individual is a function only of θ and of ω. Thus, conditional on drawing an $\omega \leq \omega^*$, any decision-maker's best response is the same under F and F' (for a given θ). Since no individual lies downwards, all states $\omega > \omega^*$ will be reported as some $r > r^* = \omega^*$. Thus, the distribution of reports for $\hat{r} \leq r^* = \omega^*$ must be the same. g-invariance and o-invariance follows from the same reasoning as before. Full support follows so long as some individuals have sufficiently large lying costs. Uniqueness of the reporting distribution follows for the same reasoning as in the previous observation. □

C.2 Reputation for Honesty (RH)

Observation 19 *Suppose individuals have RH utility. We observe drawing in and g-invariance. Some individuals will lie downwards in equilibrium when states are unobserved but never with observability (and so we observe o-shift). Last, an equilibrium with full support can only exist if $f(\omega_0) > 0.5$.*

Proof: We suppose an equilibrium with full support exists. From the observation regarding behavior in the Generalized Reputation model (which nests the RH model), presented in Section B.3, we know that in any equilibrium with full support there will be a single type of θ who is indifferent between r_0 and r_1, and that this does not depend on ω. Thus there will be downwards lying when the state is unobserved.

It follows that $\Lambda(r_1) = \frac{\text{Prob}(\theta \leq \bar{\theta}) f(\omega_0)}{\text{Prob}(\theta \leq \bar{\theta}) f(\omega_0) + \text{Prob}(\theta \leq \bar{\theta})[1 - f(\omega_0)]} = f(\omega_0)$. Thus the probability of a report of r_1 being made by a liar is equal to the probability of having drawn ω_0. Similarly,

$\Lambda(r_0) = \frac{\text{Prob}(\theta>\bar{\theta})f(\omega_1)}{\text{Prob}(\theta>\bar{\theta})f(\omega_1)+\text{Prob}(\theta>\bar{\theta})[1-f(\omega_1)]} = f(\omega_1) = 1 - f(\omega_0)$, i.e., the probability of a report of r_0 being made by a liar is equal to the probability of having drawn ω_1. Importantly, although the probability of being a liar depends on the distribution of reports, because, in equilibrium, reports do not depend on the drawn state, it is the case that the probability of being a liar in actuality only depends on the distribution F, and not on H or G. Thus, the indifferent type will not change with changes in H or G.

For the type $\bar{\theta}$ who is indifferent between the two reports it then must be the case that $\phi(r_0, \text{Probability}[\omega_i \neq 0 | r_i = 0]; \bar{\theta}) = \phi(1, \text{Probability}[\omega_i \neq 1 | r_i = 1]; \bar{\theta})$. This equality reduces to $\phi(r_0, 1 - f(\omega_0); \bar{\theta}) = \phi(1, f(\omega_0); \bar{\theta})$. There will be a single type that is indifferent because of the assumption on the cross partial of Λ and θ. And because there must be only a single indifferent type the equilibrium is unique. An equilibrium will exist given the continuity of ϕ and the fact that Λ is continuous in the cutoff $\bar{\theta}$ (although it may be a corner equilibrium without full support on all reports).

Thus, if $f(\omega_0) \leq 1 - f(\omega_0)$ then the equilibrium will not have full support, i.e., not all reports occur with positive probability, since $\phi(r_0, 1 - f(\omega_0), \bar{\theta}) < \phi(r_1, 1 - f(\omega_0), \bar{\theta}) < \phi(r_1, f(\omega_0); \bar{\theta})$ for any possible threshold.

To see that we observe g-invariance, consider a decision-maker with $\theta = \bar{\theta}$ who is exactly indifferent between reporting r_0 and r_1. Then we know this is true if and only if $\phi(r_0, 1 - f(\omega_0); \bar{\theta}) = \phi(r_1, f(\omega_0); \bar{\theta})$. Since, as described, this indifference condition varies with F, but not with beliefs about the observed reports, we observe g-invariance. Moreover, observe that if $f(\omega_1)$ increases, then the LHS of the indifference condition rises, and the RHS falls. Thus, the indifferent type must rise by the assumption on the cross partial of Λ and θ. This immediately implies we observe drawing in when moving from an equilibrium with full support.

If the state as well as the report is known then no inference needs to be done, the probability of being a liar is either 0 or 1, and so behavior should change. In particular, because it is now known whether an individual is lying or not for sure, people with low θs will give the highest report, and those with high θs will say the truth. Thus we will not have lying down. □

We can generalize part of the previous observation to the case when there are many states and many reports. Suppose there are M possible states ω, and an equal number of corresponding actions r. Observe that, as with only 2 states and actions, utilities are the same

for all individuals of type θ for any given action, regardless of the state. Thus, conditional on θ the best action r of any individual does not vary with ω. This allows us to observe the following fact: if a state that is higher has a (weakly) higher probability of being drawn, then no one will report the lower state. This says that an equilibrium with full support in G cannot exist when there is a pair of states where the higher state has a weakly higher chance of being drawn. This is the case, for example, if F is uniform.

Observation 20 *Under RH if there is an $\omega < \omega'$ such that $f(\omega) \leq f(\omega')$ then no full support equilibria exist.*

Proof: Suppose a full support equilibria exists. In this, denote $\Theta_{r=\omega}$as the set of types willing to report r. Observe that the proportion of liars at $r = \omega$ is then

$$\frac{\int_{\Theta_{r=\omega}} h(\theta)d\theta - f(\omega)\int_{\Theta_{r=\omega}} h(\theta)d\theta}{\int_{\Theta_{r=\omega}} h(\theta)d\theta} = 1 - f(\omega = r)$$

By analogous reasoning, the proportion of liars at $r' = \omega'$ is $1-f(\omega' = r')$. This means that the proportion of liars is smaller at r'. Thus the reputation cost is lower, and the monetary payoff is higher, so no one will report r. □

C.3 Conformity in Lying Costs (CLC)

Observation 21 *Suppose individuals have CLC utility. With a unique equilibrium we observe drawing out and we always observe affinity. The observability of the state does not affect the set of equilibria. Individuals never lie downwards. Equilibria with full support can exist for any F where $f(\omega_0) \in (0, 1)$.*

Proof: First, observe no individuals will lie down, because they would pay a weakly higher lying cost and receive a lower r than if they told the truth. Moreover, so long as some individuals experience sufficiently high lying costs, they will not lie (as in the LC model), and so the equilibrium will feature full support for any F where $f(\omega_0) \in (0, 1)$.

Next, consider an equilibrium with full support. The indifferent types between the two reports that drew ω_0 satisfy: $\phi(r_0, \eta(0, \bar{c}); \bar{\theta}) = \phi(r_1, \eta(c, \bar{c}); \bar{\theta})$ where c denotes the cost of lying to report r_1. The threshold type may not be unique, and so the reporting distribution

may not be unique.[53] An equilibrium will exist given the continuity of ϕ and η and the fact that $\bar{c}$ is continuous in the cutoff $\bar{\theta}$ (although it may be a corner equilibrium without full support on all reports).

However, for the moment, suppose that the equilibrium is unique in order to consider what happens when $f(\omega_0)$ falls. Consider the function $\zeta(\bar{\theta})$, which maps from Θ to Θ. Given a threshold $\bar{\theta}$ that all other individuals are using, $\zeta(\bar{\theta})$ gives the optimal threshold for any given individual. Observe that since equilibrium behavior is characterized by the threshold $\bar{\theta}$, finding the fixed point(s) of $\zeta(\bar{\theta})$ characterizes the equilibrium (this is true even for multiple equilibria). Observe that for any $\bar{\theta}$ as $f(\omega_0)$ falls $\bar{c}$ must fall. Thus $\zeta(\bar{\theta})$ must fall for all $\bar{\theta}$ (again this is true even with multiple equilibria). Thus the fixed point (which we supposed was unique) must fall. Intuitively, the indifferent type must fall as well since lying becomes more costly. So fewer people who draw ω_0 report r_1. Thus we observe drawing out. However, if we relax the assumption of equilibrium uniqueness, we cannot make a specific prediction regarding drawing in or drawing out.

Since G enters in the the utility function directly (because G has a one-to-one mapping with $\bar{r}$) we can still make predictions regarding the effect of $\hat{G}$ even though the reporting distribution may not be unique. To see that we observe affinity, notice that fixing F, increasing $\hat{g}(r_1)$ implies that the individual believes that there are more liars. Thus the costs of lying fall, and so more individuals are willing to lie.

As with the LC model, our interpretation of these costs as internal costs means that they do not depend on the inference others are making about any given person. Thus, individuals do not care whether their state was observed. Thus the set of possible equilibria is not affected by observability of the true state, and the prediction regarding lying downwards is the same for observable or unobservable states. □

C.4 Separable Reputation for Honesty plus Lying Costs (RH+LC)

Observation 22 *Suppose individuals have Separable RH+LC utility. With a unique equilibrium we observe drawing in. Individuals may lie downwards when the state is unobserved, but will not when states are observed (and so we observe o-shift). Equilibria with full support can*

[53] Intuitively this is true because if others are more likely to lie, I am more likely to lie, and so our lying behaviors are complements.

exist for any F where $f(\omega_0) \in (0,1)$.

Proof: Since the Separable RH+LC model is a special case of the General RH+LC model, we will refer to the proof of the general model (Appendix B.15) when those results carry over. As in the general model, with sufficiently high lying costs, the model can always generate a distribution of reports with full support for any F where $f(\omega_0) \in (0,1)$.

Similarly, as in the general RH+LC model, there is a threshold function for each state $\tau_{\omega_i}(\theta_1, \theta_2)$. Unlike in the general model, we can characterize the equilibrium in terms of the intercepts of the threshold function, rather than the probability of being a liar. To see this, observe that the threshold functions are always linear in θ_1 and θ_2 since the utility function itself is linear in those parameters. Given a particular utility function, this line can be characterized by its θ_1 intercept and its θ_2 intercept denoted $\theta_{1T}^{\omega_i}$ and $\theta_{2T}^{\omega_i}$. Moreover, since the LC portion of costs never depends on the distribution of responses, the $\theta_{1T}^{\omega_i}$ intercept (i.e. the threshold value of $\theta_{1T}^{\omega_i}$ when $\theta_2 = 0$) must always be the same, and so in fact, we can think of this line as really being characterized by a single intercept: $\theta_{2T}^{\omega_i}$. The thresholds $\theta_{2T}^{\omega_i}$ (one for each state), along with H, induce a conditional (on each state) probability of giving either report. These, in conjunction with F, define the probability of being a liar at either report (as well as G). Thus, in order to solve for an equilibrium we can consider a function $\zeta(\theta_{2T}^{\omega_1}, \theta_{2T}^{\omega_2})$ which maps from the thresholds that everyone is using into best response thresholds.

However, observe that because we are looking at the θ_2 intercepts, this is where the standard lying costs are 0. Thus, the actual drawn state does not enter the utility function, and so players must behave the same regardless of which state they drew; so $\theta_{2T}^{\omega_1} = \theta_{2T}^{\omega_2}$. Thus, our problem reduces to a single dimension; and we can consider function $\zeta(\theta_{2T})$ and find its fixed point. An equilibrium will exist given the continuity of ϕ and that fact that Λ is continuous in the threshold sets (although it may be a corner equilibrium without full support on all reports).

Now consider the threshold θ_{2T}. It is defined as the solution to the equation $u(r_1) - \tilde{\theta}_2 \upsilon(\Lambda(r_1)) = u(r_0) - \tilde{\theta}_2 \upsilon(\Lambda(r_0))$ or $u(r_1) - u(r_0) = \tilde{\theta}_2(\upsilon(\Lambda(r_1)) - \upsilon(\Lambda(r_0)))$. This describes an individual with $\theta_1 = 0$ and a $\theta_2 = \tilde{\theta}_2$ such that the individual is indifferent between reporting r_0 or r_1. If $\tilde{\theta}_2 = 0$ (which implies everyone says r_0) the RHS of this equation is equal to 0. Thus, a sufficient condition for a unique equilibrium is if the RHS is monotone increasing in

$\tilde{\theta}_2$. Unfortunately we cannot guarantee this. As $\tilde{\theta}_2$ increases, the probability, conditional on drawing ω_0, of reporting r_1 increases. Similarly, the probability, conditional on drawing ω_1, of reporting r_1 increases. Thus, at r_1 (and similarly r_0) there are both more truth-tellers and more liars, making the change in $\Lambda(r_1)$, and the change in the difference $\upsilon(\Lambda(r_1)) - \upsilon(\Lambda(r_0)))$ ambiguous.

Condition on a unique equilibria and suppose that $f(\omega_1)$ increases. Then, fixing the threshold types (i.e. strategies) of all others, the proportion of truth-tellers must increase at r_1. Similarly, the proportion of truth-tellers at r_0 must fall. This makes r_1 relatively more attractive to individuals (compared to r_0). Thus the optimal threshold θ_2 (generated by ζ) must rise and we get drawing in.

The other statements follow from the General RH+LC model. □

We have no prediction regarding our $\hat{G}$ treatment for the same reason as in the general RH+LC model.

C.5 Details of the CLC Calibration

This section describes the details of the calibration of the CLC model presented in Section 4. We calibrate the CLC model in order to understand the potential size of the $\hat{G}$ treatment effect. For the calibration, we make a number of assumptions. First, we assume that utility takes the form $r - \theta\kappa\frac{c(r,\omega)}{\bar{c}}$. κ indicates the non-idiosyncratic strength of the cost of lying. $c(r,\omega)$ takes on the value 0 if $r = \omega$, and 1 if $r \neq \omega$. Recall $\bar{c}$ is the average cost of lying in society, and so here is equivalent to the fraction of liars. Moreover, since no individuals lie down in the CLC model, this simply represents the fraction of people who drew ω_0 but report r_1. We normalize $r_0 = -1$ and $r_1 = 1$. Moreover, we will suppose that θ is uniformly distributed on $[0, 1]$. Given an equilibrium with full support the threshold type (who draws the low state) must satisfy the condition $1 - \bar{\theta}\kappa\frac{1}{\bar{c}} = -1$ or $\bar{\theta} = \frac{2\bar{c}}{\kappa}$. We can calibrate the threshold by observing that the proportion of high reports was 0.45 in the F_LOW treatment, and so 35 percent of the population lied. This indicates $\bar{\theta} = \frac{0.7}{\kappa}$. Moreover, given the uniform distribution of θ, this also implies $\bar{\theta} = \frac{0.35}{0.9} = 0.389$, so $\kappa = 1.8$. Given this, suppose that $f(\omega_1) = 0.1$ and that $\bar{c}$ shifts from 0.31 to 0.52 which is the shift implied by the average change in beliefs in our $\hat{G}$ treatment, since our treatment shifted beliefs about the proportion of high

reports from 0.41 to 0.62. Then the threshold type shifts from $\frac{0.62}{1.8} = 0.344$ to $\frac{1.08}{1.8} = 0.578$, implying that 21 percent of subjects (since 90 percent of subjects draw the low state) will increase their report across treatments.

More broadly, if social comparison models are calibrated so as to fit other facets of our data (i.e., full support or drawing in), social comparisons must be a reasonably large component of utility. Given this, and the assumption that the marginal types (and types close to them) are drawn with "reasonable" frequency, it must be the case that a relatively large fraction of subjects should respond to shifts in beliefs about G.

C.6 Details of the RH+LC Calibration

This section describes the details of the calibration of the RH+LC model presented in Section 4. We first prove the observation in the text.

Observation 1: *The set of equilibria implied by the utility function (1) is invariant to affine shifts in monetary payoffs , i.e., if the possible states ω and reports r are replaced by $\alpha\omega + \beta$ and $\alpha r + \beta$, with $\alpha, \beta \in \mathbb{R}$ and $\alpha > 0$.*

Proof: Suppose that, conditional on drawing ω an individual is indifferent between reporting r reporting $\hat{r}$. Then we know that $r - \kappa_1\theta_1|r-\omega| - \kappa_2\theta_2\Lambda(r)\frac{(r_{max}-r_{min})}{2} = \hat{r} - \kappa_1\theta_1|\hat{r}-\omega| - \kappa_2\theta_2\Lambda(\hat{r})\frac{(r_{max}-r_{min})}{2}$. Now consider what happens when we multiply all monetary payoffs by α and add β. The indifference condition is now $\alpha r + \beta - \kappa_1\theta_1|\alpha r + \beta - \alpha\omega - \beta| - \kappa_2\theta_2\Lambda(\alpha r + \beta)\frac{(\alpha r_{max}+\beta-\alpha r_{min}-\beta)}{2} = \alpha\hat{r} + \beta - \kappa_1\theta_1|\alpha\hat{r} + \beta - \alpha\omega - \beta| - \kappa_2\theta_2\Lambda(\alpha\hat{r} + \beta)\frac{(\alpha r_{max}+\beta-\alpha r_{min}-\beta)}{2}$ or $\alpha r - \alpha\kappa_1\theta_1|r-\omega| - \alpha\kappa_2\theta_2\Lambda(\alpha r+\beta)\frac{(r_{max}-r_{min})}{2} = \alpha\hat{r} - \alpha\kappa_1\theta_1|\hat{r}-\omega| - \alpha\kappa_2\theta_2\Lambda(\alpha\hat{r}+\beta)\frac{(r_{max}-r_{min})}{2}$, which holds if and only if: $r - \kappa_1\theta_1|r-\omega| - \kappa_2\theta_2\Lambda(\alpha r + \beta)\frac{(r_{max}-r_{min})}{2} = \hat{r} - \kappa_1\theta_1|\hat{r}-\omega| - \kappa_2\theta_2\Lambda(\alpha\hat{r} + \beta)\frac{(r_{max}-r_{min})}{2}$. Observe that $\Lambda(r)$ does not depend on the actual numerical value of the report, but only on the rank of the report within the possible set of reports. Thus the indifference conditions are the same when we multiply all monetary payoffs by α and add β.□

In order to calibrate the model we focus on a trinary state and report model with a uniform F and a uniform joint distribution of the θs on $[0,1] \times [0,1]$. We normalize monetary payoffs to -1, 0 and 1. Moreover, we will focus on an equilibrium where no individuals lie downwards, so that individuals who draw 1 will report 1. In contrast, conditional on drawing -1, an individual must decide whether to report -1, 0 or 1 (here we will simply denote the states

and the reports using the normalized payoff associated with each one). The threshold set for this individual to be indifferent between reporting -1 and 0 satisfies $0 = 1 - \kappa_1\theta_1 - \kappa_2\theta_2\Lambda(0)$. The threshold set for this individual to be indifferent between reporting -1 and 1 satisfies $0 = 2 - 2\kappa_1\theta_1 - \kappa_2\theta_2\Lambda(1)$. Last, the threshold for this person to be indifferent between reporting 0 and 1 satisfies $0 = 1 - \kappa_1\theta_1 - \kappa_2\theta_2(\Lambda(1) - \Lambda(0))$.

Recall that we focus on a model where someone drawing -1 is choosing between reporting one of the three values. Thus, only the first and third threshold conditions are relevant. Again, we will verify that the actual realized values in the calibration are consistent with this assumption.

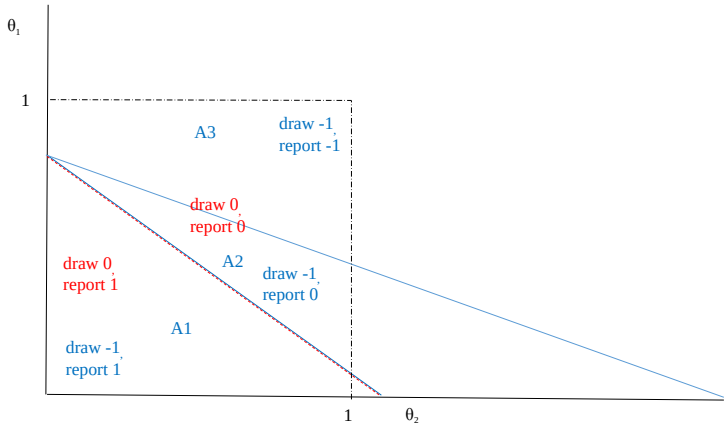


Figure C.1: Thresholds for RH plus LC Model: Trinary States/Reports

Conditional on drawing 0, an individual must decide whether to report 0 or 1 (recall no one lies down), and the threshold types satisfy the equation

$$0 = 1 - \kappa_1\bar{\theta}_1 - \kappa_2\bar{\theta}_2(\Lambda(1) - \Lambda(0))$$

Observe that this is the same equation that the individuals drawing -1 faces when deciding between 0 and 1. Figure C.1 demonstrates graphically the thresholds and reporting strategies for individuals depending on the draw and their θs. $A1$, $A2$ and $A3$ represent the areas of the respective regions. These regions are defined in terms of the κs , $\Lambda(1)$ and $\Lambda(0)$ using the threshold functions discussed above.

Using the areas, the probability of being a liar at 1 is then: $\Lambda(1) = \frac{\frac{1}{3}A1+\frac{1}{3}A1}{\frac{1}{3}+\frac{1}{3}A1+\frac{1}{3}A1}$. Similarly, $\Lambda(0) = \frac{\frac{1}{3}A2}{\frac{1}{3}(1-A1)+\frac{1}{3}A2}$. Given any values of κ_1 and κ_2 we now have a pair of equations with two unknowns: $\Lambda(1)$ and $\Lambda(0)$. Moreover $g(-1) = \frac{1}{3}(1-A1-A2) = \frac{1}{3}A3$, $g(0) = \frac{1}{3}(1-A1)+\frac{1}{3}A2$ and $g(1) = \frac{1}{3}+\frac{1}{3}A1+\frac{1}{3}A1$. With $\kappa_1 = 3$ and $\kappa_2 = 4$ we obtain $\Lambda(1) = .286$ and $\Lambda(0) = .0863$, and $g(-1) = .241$, $g(0) = .292$, $g(1) = .467$.

Algebra verifies that individuals will not lie down. For an individual drawing 1, the value of reporting 1 is $1 - 1.144\theta$; the value of reporting 0 is $-2 - .344\theta$; and the value of reporting -1 is -5. It is easy to verify that reporting 1 is the best action. For an individual drawing 0 the value of reporting 0 is $1 - .344\theta$. The value of reporting -1 is -2. Again, it is clear that reporting 0 is the better action. Moreover, we also find that some individuals draw -1 and report 0, again consistent with our assumption, since $\Lambda(0) \neq 0$.

We can also use this calibrated utility function to consider what should happen in our experimental setting; again assuming a uniform distribution over possible idiosyncratic cost parameters. Now there are only two normalized states; -1 and 1. Conditional on drawing -1, an individual must decide whether to report -1 or 1. Thus, if there are downwards liars (which there may be, given the parameterization and our potential F distributions), the indifferent type satisfies

$$-1 - \kappa_2\bar{\theta}_2\Lambda(-1) = 1 - 2\kappa_1\bar{\theta}_1 - \kappa_2\bar{\theta}_2\Lambda(1)$$

Thus, the threshold function is $\bar{\theta}_1 = \frac{1}{\kappa_1} - \frac{\kappa_2}{\kappa_1}\frac{\bar{\theta}_2}{2}(\Lambda(1) - \Lambda(-1))$. We can denote the area below the threshold within the unit square as B_0 and the the area above a B_1. Similarly the indifferent type who drew the high state has a threshold function $\bar{\theta}_1 = -\frac{1}{\kappa_1} + \frac{\kappa_2}{\kappa_1}\frac{\bar{\theta}_2}{2}(\Lambda(1) - \Lambda(-1))$. We can denote the area below this threshold within the unit square as B_3 and above as B_2. See Figure C.2 for a graphical representation of these areas (which also show's the threshold functions (i.e. the τs).

Using the areas, the probability of being a liar at 1 is then: $\Lambda(1) = \frac{f(-1)B_0}{f(1)B_2+f(-1)B_0}$. Similarly, $\Lambda(0) = \frac{f(1)B_3}{f(1)B_3+f(-1)B_1}$. Given any values of κ_1 and κ_2 we now have a pair of equations with two unknowns: $\Lambda(1)$ and $\Lambda(0)$. Moreover, $g(1) = f(1)B_2 + f(-1)B_0$ and $g(0) = f(1)B_3 + f(-1)B_1$. We can then substitute using F under our two treatments (recall $f(-1) = 0.9$ in F_LOW and $f(-1) = 0.4$ in F_HIGH). Solving for the solutions of our

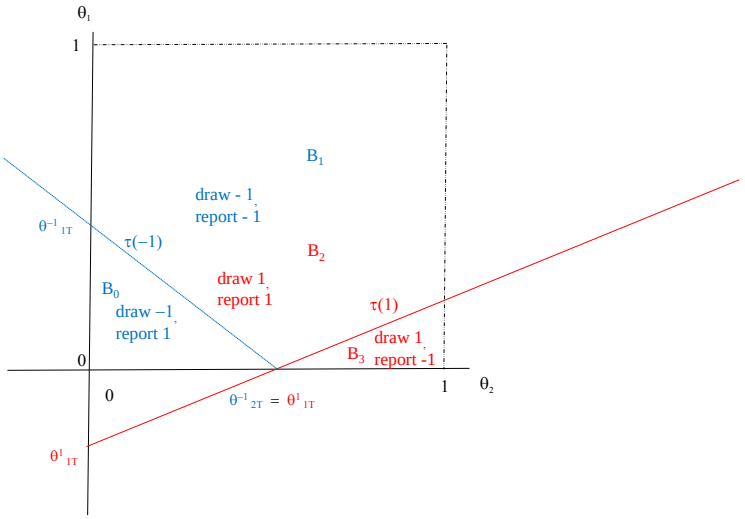


Figure C.2: Thresholds for RH plus LC Model: Binary States/Reports

equations then gives that under F_LOW $\Lambda(1) = 0.569, \Lambda(-1) = .00036, g(-1) = 0.768, g(1) = 0.232$ and $1 - \frac{g(-1)}{f(-1)} = 0.146$; while under F_HIGH $\Lambda(1) = 0.158, \Lambda(-1) = 0, g(-1) = 0.289, g(1) = 0.712$, and $1 - \frac{g(-1)}{f(-1)} = 0.277$. Moreover, these are both unique equilibria.

Clearly, our calibrated model underpredicts lying compared to our actual experimental data. This should not be a surprise, as we observe a higher average standardized report in our experiment than we do in our meta-analysis. However, we still observe quite a strong drawing in effect in our calibrated model.

Moreover, because the calibration implies that there will be lying downwards when $f(\omega_0) = 0.9$, we can get anything between aversion and affinity (depending on how individuals update their belief about the joint distribution of the θs using $\hat{G}$).

D Experimental Instructions

In this appendix we reproduce the instructions used in our experiments. We first present the instructions and questionnaire used in the F_LOW treatment and highlight, using numbers in square brackets, where and how the F_HIGH treatment instructions differ. We then present the instructions for the G_LOW treatment and highlight the differences for G_HIGH. Then we present the instructions for the OBSERVABLE and UNOBSERVABLE treatments. Finally, a photo of the lab setup.

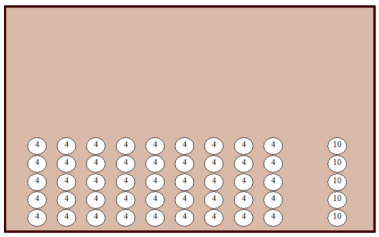
D.1 Instructions for F_LOW

Welcome and thank you for taking part in this study. The study is run by the "Centre for Decision Research and Experimental Economics" and has been financed by various research foundations. During the experiment, we request that you turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.

In this study we ask you to complete a questionnaire, which you can find on your desk.

For completing the questionnaire you will receive an amount of money. To determine how much money you will be paid, we ask you to randomly draw a chip from an envelope after completing the questionnaire. The envelope will contain 50 chips with the numbers 4 or 10 on them. The composition of the envelope is shown in the figure below. You can also see the chips on the tray in front of you.

[1]



[2] The envelope will contain 45 chips with the number 4; and 5 chips with the number 10.

The number represents the amount of money that you will be paid for this study if you draw a chip with that number. If you draw a chip with a 4, you will be paid £4; if you draw a chip with a 10, you will be paid £10. This payment already includes your show-up fee.

When you have finished the questionnaire, click the CONTINUE button that will appear on your computer screen. On the next screen you will find instructions for how to perform the

draw that will determine your payment.
After the draw, turn off your computer using the power button. Write down the number of your chip on the PAYMENT SHEET that is on your desk. Then bring the questionnaire and payment sheet to the experimenter who will be waiting outside the lab.
If you have a question now, please raise your hand and the experimenter will come to your desk to answer it.

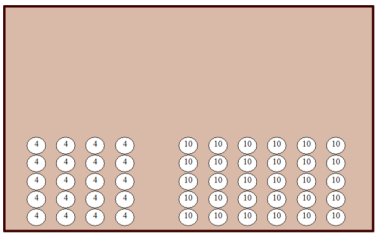
The on-screen instructions about how to perform the draw were as follows:
When you have finished your questionnaire click the CONTINUE button to proceed to the next screen where you will find instructions for how to perform the draw that will determine your payment.

On your desk you find a tray containing 50 chips with the numbers 4 or 10 on them.
Place all the chips into the brown envelope that is also placed on your desk. Shake the envelope a few times and then, without looking, randomly draw a chip from the envelope.
Your payment in £ is equal to the number of the chip you have drawn from the envelope.
After observing the outcome of the draw, place the chip back into the envelope.
When you have finished click the OK button to proceed to the next screen.

Please now turn off your computer using the power button and write down the number of your chip on your payment sheet.
Then bring the questionnaire and the payment sheet to the experimenter who is waiting outside.

D.2 Instructions for F_HIGH

The instructions for F_HIGH are identical to the ones for F_LOW except in two places:

[1] 

[2] The envelope will contain 20 chips with the number 4; and 30 chips with the number 10.

D.3 Questionnaire used in the F_LOW and F_HIGH experiments

QUESTIONNAIRE

This is a questionnaire consisting of 22 questions.

Please complete this questionnaire as clearly and accurately as possible. All your responses will be completely confidential. Please leave blank any questions you do not feel comfortable answering.

Thank you in advance for your cooperation.

QUESTIONS

1. What is your gender? Answ: Female Male

2. What is your age? Answ: _____ years

3. What is your nationality? (Open answer)

4. Are you currently: Married; Living together as married; Separated; Widowed; Single

5. What is your major area of study? Answ: Engineering; Economics; Law; Business economics; Political economics; Other Social sciences; Humanities; Health-related sciences; Natural sciences; Other (please specify) _____

6. Which of the following ethnic groups is appropriate to indicate your cultural background? Answ: White; Mixed; Asian or Asian British; Black or Black British; Chinese; Other ethnic group (please specify) _____

7. How important is religion to you? Answ: Very important; Moderately important; Mildly important; Not important

8. How would you rate your money management? (the way you handle your finances) Answ: Poor; Average; Good; Excellent

9. How would you rate your knowledge of financial products such as ISAs, credit cards, loans and mortgages? Answ: Poor; Average; Good; Excellent

10. Whilst growing up, were your parents/guardians open to discussing financial matters within the home? Answ: YES NO

11. Since becoming a student & receiving maintenance loans/grants, would you say that you budget effectively or that you struggle to purchase basic necessities? (Necessities meaning food, toiletries and standard living costs - not eating out) Answ: I've always known how to budget; I've had to learn to budget whilst at University; I struggle to purchase necessities; I can afford everything but I don't budget

12. If you struggle to purchase necessities, what would you put this down to? Answ: Not budgeting; Cost of necessities too expensive; Too care-free with money; Other priorities such as shopping & nightlife take a priority; I don't struggle, I'm good with budgeting; I have no idea

13. What are your top five spending priorities? (Open Answer)

14. Do you regularly know how much money you have in your bank account? Answ: YES NO

15. Do you keep track of your spending? Answ: YES NO

16. Do you have money set aside for an emergency? Answ: YES NO

17. Are you in debt? Answ: YES NO

18. Do you shop around to get the best deal when selecting financial products such as insurance and mobile phones? Answ: YES NO

19. Do you have a job to provide extra income whilst at University? Answ: YES NO

20. If you needed financial advice tomorrow, who would you turn to? Answ: Student Union; Parents; Friends; Bank; Financial adviser; Other (please specify) _____

21. What benefits would you expect from being able to better manage your money? (Open Answer)

22. Is there anything which would help you to better manage your money? (Open Answer)

***

Thank you for completing this questionnaire.

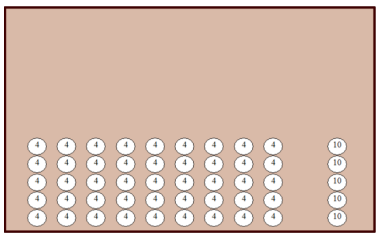
Please now follow the instructions on your computer screen to determine your payment for completing the questionnaire.

D.4 Instructions for G_LOW

Welcome and thank you for taking part in this study.
The study is run by the "Centre for Decision Research and Experimental Economics" and has been financed by various research foundations. During the study, we request that you turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.
In this study we ask you to read a short description of a potential experiment, to imagine two possible outcomes of this experiment, and then to answer several questions about these outcomes.
Please turn over to the next page to read the description of the experiment.

DESCRIPTION OF THE POTENTIAL EXPERIMENT

Imagine that we would run the following experiment here in this lab with 50 participants. Participants are invited by email and asked to complete a questionnaire about their money management. Participants receive an amount of money for completing the questionnaire. To determine how much money they are paid, participants are asked to randomly draw a chip from an envelope after completing the questionnaire. The envelope contains 50 chips with the numbers 4 or 10 on them. The composition of the envelope is shown in the figure below. You can also see the chips on the tray in front of you.

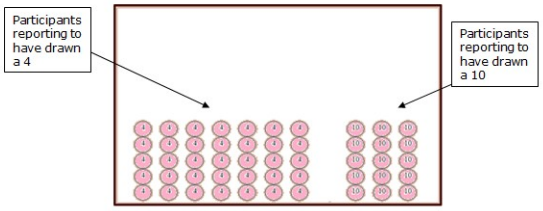


The envelope contains 45 chips with the number 4; and 5 chips with the number 10. The number represents the amount of money that participants are paid if they draw a chip with that number. If they draw a chip with a 4, they are paid £4; if they draw a chip with a 10, they are paid £10. When they have finished the questionnaire, participants are asked to place all the chips into a brown envelope that is placed on their desk (you can also see this envelope on your desk). Participants are asked to shake the envelope a few times and then, without looking, randomly draw a chip from the envelope. They are told to place the chip back into the envelope after having observed the outcome of the draw, and to write down the number of their chip on a PAYMENT SHEET and then bring the questionnaire and payment sheet to the experimenter who has been waiting outside the lab for the whole time.

POSSIBLE OUTCOMES OF THE EXPERIMENT

Imagine that the results of the experiment would be as follows.

[3] Of the 50 participants, 35 report to have drawn a 4 and 15 report to have drawn a 10. This outcome is shown in the figure below.



[4]

We now ask you to answer a number of questions about this imagined outcome of the experiment. Please consider each question carefully before answering it.

1) Each participant had an envelope containing 45 chips with the number 4 and 5 chips with the number 10. What are the chances that a participant randomly draws a 10 from the envelope? Your answer: _____

[5] 2) 15 out of 50 participants reported to have drawn a 10. How many of the participants who have reported to have drawn 10 do you think have truly drawn a 10? Your answer: _____

3) Some of the participants who drew a 4 actually reported 10. Can you imagine why they would do that? Your answer:_____

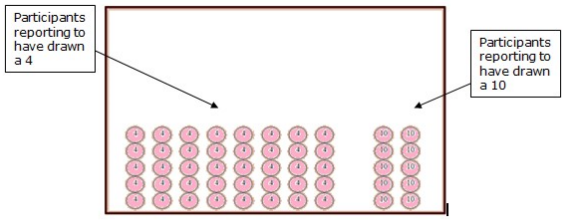
4) Some of the participants who drew a 4 actually reported 4. Can you imagine why they would do that? Your answer:_____

5) How satisfied do you think that the participants who reported a 4 will be? Your answer: very dissatisfied _____ very satisfied

6) How satisfied do you think that the participants who reported a 10 will be? Your answer: very dissatisfied _____ very satisfied

Now imagine that the results of the experiment would be as follows.

[6] Of the 50 participants, 40 report to have drawn a 4 and 10 report to have drawn a 10. This outcome is shown in the figure below.



[7]

[8] 7) 10 out of 50 participants reported to have drawn a 10. How many of the participants who have reported to have drawn 10 do you think have truly drawn a 10? Your answer: _____

8) How satisfied do you think that the participants who reported a 4 will be? Your answer: very dissatisfied _____ very satisfied

9) How satisfied do you think that the participants who reported a 10 will be? Your answer: very dissatisfied _____ very satisfied

[9] 10) Which of the two imagined outcomes described above do you think is more realistic? Your answer: The outcome where 15 out of 50 participants reported a 10; The outcome where 10 out of 50 participants reported a 10

Last year we actually ran the experiment that we just described to you here in this lab.

Please estimate the fraction (in percent) of participants in the previous experiment who reported to have drawn a 10. If your estimate is accurate with an error of at most +/- 3 percentage points we will pay you £3 at the end of this experiment.
Your answer: _____ out of 100

SOME QUESTIONS ABOUT YOURSELF

1. What is your gender? Female Male
2. What is your age? _____ years
3. What is your nationality? _____
4. What is your major area of study? Engineering; Economics; Law; Business economics; Political economics; Other Social sciences; Humanities; Health-related sciences; Natural sciences; Other (please specify) _____

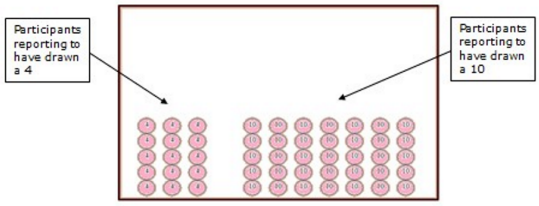
YOUR PAYMENT FOR TAKING PART IN TODAY'S STUDY

On top of the money that you may earn if you have answered the question above correctly, we will pay you an additional sum of money for having taken part in this study.
To determine how much money you will be paid we ask you to randomly draw a chip from an envelope, as the participants in the experiment that we described before. Please place all the chips that are displayed in the tray in front of you into the brown envelope that is placed on your desk. The envelope will thus contain 45 chips with the number 4 and 5 chips with the number 10. Shake the envelope a few times and then, without looking, randomly draw a chip from the envelope. Your payment in GBP is equal to the number of the chip you have drawn from the envelope.
After observing the outcome of the draw, place the chip back into the envelope. Write down the number of your chip on the first page of this document. Then bring the document to the experimenter who is waiting outside the lab.

D.5 Instructions for G_HIGH

The instructions for G_HIGH are identical to the ones for G_LOW except in the following places:

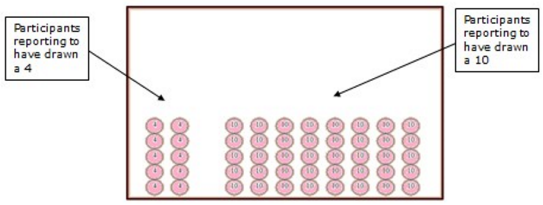
[3] Of the 50 participants, 15 report to have drawn a 4 and 35 report to have drawn a 10. This outcome is shown in the figure below.



[4]

[5] 2) 35 out of 50 participants reported to have drawn a 10. How many of the participants who have reported to have drawn 10 do you think have truly drawn a 10?

[6] Of the 50 participants, 10 report to have drawn a 4 and 40 report to have drawn a 10. This outcome is shown in the figure below.



[7]

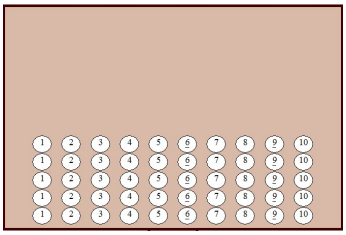
[8] 7) 40 out of 50 participants reported to have drawn a 10. How many of the participants who have reported to have drawn 10 do you think have truly drawn a 10?

[9] 10) Which of the two imagined outcomes described above do you think is more realistic? Your answer: The outcome where 35 out of 50 participants reported a 10; The outcome where 40 out of 50 participants reported a 10

D.6 Instructions for UNOBSERVABLE

Welcome and thank you for taking part in this study. The study is run by the "Centre for Decision Research and Experimental Economics" and has been financed by various research foundations. During the experiment, we request that you turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.

In this study we ask you to complete a questionnaire, which you can find on your desk.
For completing the questionnaire you will receive an amount of money. To determine how much money you will be paid, we ask you to randomly draw a chip from an envelope after completing the questionnaire. The envelope will contain 50 chips with the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10 on them. The composition of the envelope is shown in the figure below. You can also see the chips on the tray in front of you.



The envelope will contain exactly 5 chips of each number.
The number represents the amount of money that you will be paid for this study if you draw a chip with that number. For instance, if you draw a chip with a 1, you will be paid £1; if you draw a chip with a 2, you will be paid £2; and so on; if you draw a chip with a 10, you will be paid £10.
When you have finished the questionnaire, click the CONTINUE button that will appear on your computer screen. On the next screen you will find instructions for how to perform the draw that will determine your payment.
After the draw, turn off your computer using the power button. Write down the number of your chip on the PAYMENT SHEET that is on your desk. Then bring the questionnaire and payment sheet to the experimenter who will be waiting outside the lab.
If you have a question now, please raise your hand and the experimenter will come to your desk to answer it.

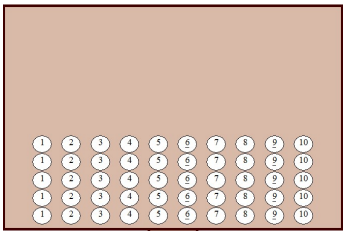
The on-screen instructions about how to perform the draw were as follows:
When you have finished your questionnaire click the CONTINUE button to proceed to the next screen where you will find instructions for how to perform the draw that will determine your payment.

On your desk you find a tray containing 50 chips with the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10 on them.
Place all the chips into the brown envelope that is also placed on your desk. Shake the envelope a few times and then, without looking, randomly draw a chip from the envelope.
Your payment in £ is equal to the number of the chip you have drawn from the envelope.
After observing the outcome of the draw, place the chip back into the envelope.
When you have finished click the OK button to proceed to the next screen.

Please now turn off your computer using the power button and write down the number of your chip on your payment sheet.
Then bring the questionnaire and the payment sheet to the experimenter who is waiting outside.

D.7 Instructions for OBSERVABLE

Welcome and thank you for taking part in this study. The study is run by the "Centre for Decision Research and Experimental Economics" and has been financed by various research foundations. During the experiment, we request that you turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.
In this study we ask you to complete a questionnaire, which you can find on your desk.
For completing the questionnaire you will receive an amount of money. To determine how much money you will be paid, we ask you to randomly draw a chip from an envelope after completing the questionnaire. The envelope will contain 50 chips with the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10 on them. The composition of the envelope is shown in the figure below.



The envelope will contain exactly 5 chips of each number.
The number represents the amount of money that you will be paid for this study if you draw a chip with that number. For instance, if you draw a chip with a 1, you will be paid £1; if

you draw a chip with a 2, you will be paid £2; and so on; if you draw a chip with a 10, you will be paid £10.

When you have finished the questionnaire, click the CONTINUE button that will appear on your computer screen. On the next screen you will find instructions for how to perform the draw that will determine your payment.

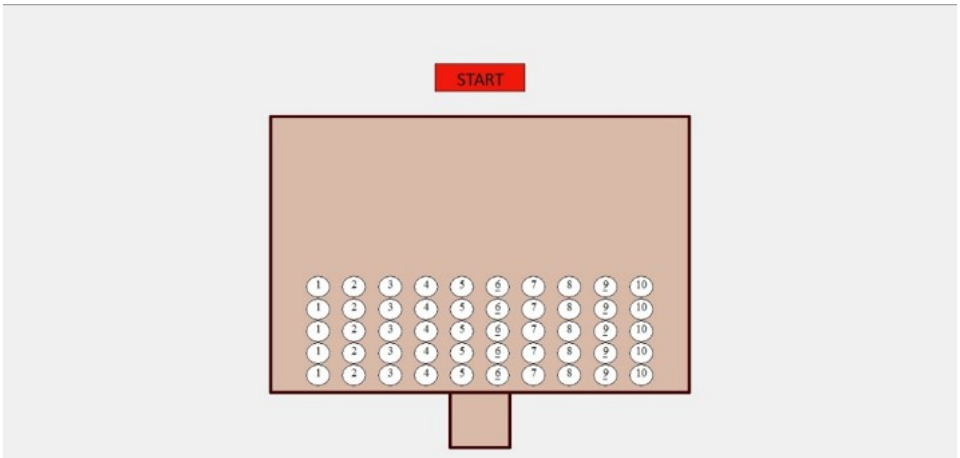
After the draw, open the brown envelope that is placed on your desk. The envelope contains 10 coins of £1 each. Take as many coins as the number of the chip you have drawn. Then turn off your computer using the power button and quietly exit the lab leaving these instructions, your questionnaire, and the brown envelope on the desk. (Note: you do not have to sign a receipt for this experiment).

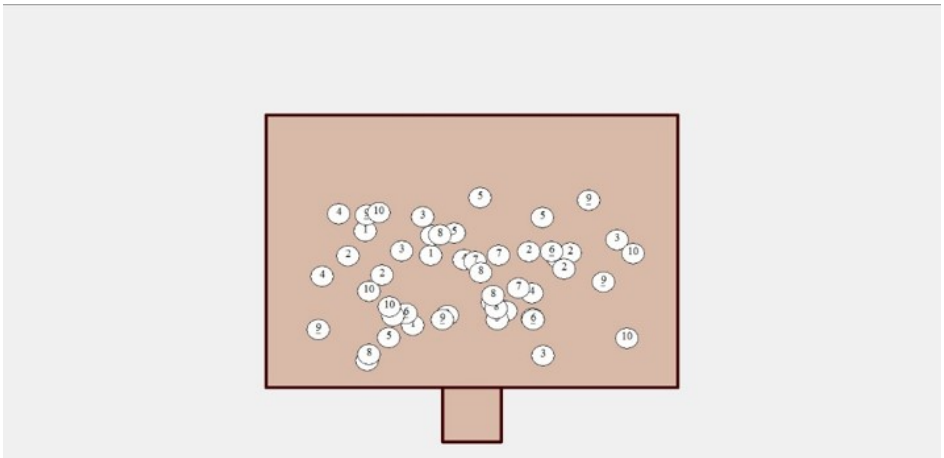
If you have a question now, please raise your hand and the experimenter will come to your desk to answer it.

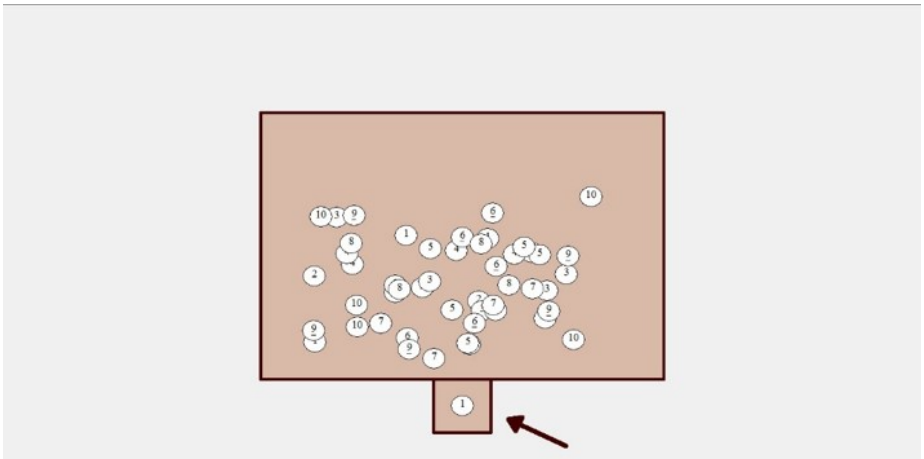
The on-screen instructions about how to perform the draw were as follows:

When you have finished your questionnaire click the CONTINUE button to proceed to the next screen where you will find instructions for how to perform the draw that will determine your payment.

Click the START button to shake the envelope. One of the chips will fall out of the envelope. Your payment in £ is equal to the number on the chip that falls out of the envelope.







Please now open the brown envelope that you can find on your desk. The envelope contains 10 coins of £1 each. Take as many coins as the number of the chip you have drawn.
Then turn off your computer using the power button (click only once and then release) and quietly leave the lab, leaving all material on your desk. (Note: you do not have to sign a receipt for this experiment.)

D.8 Laboratory setup