Preferences for Truth-telling

Johannes Abeler
Daniele Nosenzo
Collin Raymond

CESIFO WORKING PAPER NO. 6087
CATEGORY 13: BEHAVIOURAL ECONOMICS
ORIGINAL VERSION: SEPTEMBER 2016
THIS VERSION: MARCH 2019

ISSN 2364-1428

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 6087*

Preferences for Truth-telling

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 90 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. Our empirical evidence suggests that a preference for being seen as honest and a preference for being honest are the main motivations for truth-telling.

JEL-Codes: D030, D820, H260, I130, J310.

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

Johannes Abeler
University of Oxford
United Kingdom - OX1 3UQ Oxford
johannes.abeler@economics.ox.ac.uk

Daniele Nosenzo
University of Nottingham
United Kingdom - Nottingham NG7 2RD
Daniele.Nosenzo@nottingham.ac.uk

Collin Raymond
Purdue University
Lafayette / IN / USA
collinbraymond@gmail.com

21 March 2019
This is the version of the paper that was accepted for publication by Econometrica.
JA thanks the ESRC for financial support under grant ES/K001558/1. We thank Steffen Altmann, Steve Burks, Gary Charness, Vince Crawford, Martin Dufwenberg, Armin Falk, Urs Fischbacher, Simon Gächter, Philipp Gerlach, Tobias Gesche, David Gill, Uri Gneezy, Andreas Grunewald, David Huffman, Navin Kartik, Michael Kosfeld, Erin Krupka, Dmitry Lubensky, Daniel Martin, Takeshi Murooka, Simone Quercia, Heiner Schuhmacher, Klaus Schmidt, Jonathan Schulz, Daniel Seidman, 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, Leander Heldring, 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, Nathan Nunn, David Pascual Ezama, Eyal Pe'er, Marco Piovesan, Matteo Ploner, Wojtek Przepiorka, Heiko Rauhut, Tobias Regner, Rainer Rilke, Ismael Rodriguez-Lara, 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. Martin Hadley, Sunham Kim, Felix Klimm, Jeff Kong, 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 made the useful simplifying assumption that utility only depends on material payoffs. In situations of asymmetric information this implies that people are not intrinsically concerned about lying or telling the truth and, if misreporting cannot be detected, individuals should submit the report that yields the highest 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. 2014).

In this paper, we aim to deepen our understanding of how people report private information. Our strategy to do so is threefold. We first conduct a meta study of the existing experimental literature and document that behavior is indeed far from the assumption of payoff-maximizing reporting. We then formalize a wide range of explanations for this aversion to lying and show that many of these are consistent with the behavioral regularities observed in the meta study.[1] Finally, in order to distinguish among the many and varied explanations, we identify new empirical tests and implement them in new experiments.

In order to cleanly identify the motivations driving aversion to lying, we focus on a setting without strategic interactions. We thus abstract from sender-receiver games or verification of messages, such as audits. We do so because the strategic interaction makes the setting more complex, especially if one is interested in studying the underlying motives of reporting behavior, as we are. We therefore use the experimental paradigm introduced by Fischbacher and Föllmi-Heusi (2013): subjects privately observe the outcome of a random variable, report the outcome and receive a monetary payoff proportional to their report (for related methods using inferences about the population, see Batson et al. 1997 and Warner 1965). While no

[1] We will use the terms "aversion to lying" and "preference for truth-telling" interchangeably (but see Sánchez-Pagés and Vorsatz 2009).

individual report can be identified as truthful or not (and subjects should thus report the payoff maximizing outcome under the standard economic assumption), 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 2017, Cohn and Maréchal forthcoming, Cohn et al. 2015, Gächter and J. F. Schulz 2016b, Potters and Stoop 2016, Dai et al. 2018).[2]

In the first part of our paper (Section 1 and Appendix A), we combine data from 90 studies that use setups akin to Fischbacher and Föllmi-Heusi (2013), involving more than 44000 subjects across 47 countries. Our study is the first quantitative meta analysis of this experimental paradigm. Interactive versions of the analyses can be found at www.preferencesfortruthtelling.com. 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 comparable to many other widely discussed non-standard behaviors observed in laboratory experiments, like altruism or reciprocity.[3] This strong preference for truth-telling is robust to increasing the payoff level 500-fold or repeating the reporting decision up to 50 times. 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.[4]

[2] 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 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, although the added strategic complexity makes it harder to identify subjects' motivations for lying. 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 M. 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 M. Dufwenberg show that the message affects the action, and 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).

[3] Our results imply that in a typical experiment based on the Fischbacher and Föllmi-Heusi (2013) paradigm and offering a maximum payment of $1, subjects take on average only 62c home and thus forgo 38c. Altruism is often measured by the amount given in dictator-game experiments. There, subjects forgo on average 28c out of each $1 (Engel 2011). Positive reciprocity is often measured by the behavior of second-mover subjects in trust games who forgo on average 38c out of each $1 (Johnson and Mislin 2011; Cardenas and Carpenter 2008). Negative reciprocity is often measured by the behavior of second-mover subjects in ultimatum-game experiments who forgo on average less than 16c out of each $1 (Oosterbeek et al. 2004).

[4] In most experiments using this paradigm, the money obtained by reporting comes from the experimenter

In the second part of the paper (Section 2 and Appendices B, C and E), we formalize a wide range of explanations for the observed behavior, including the many explanations that have been suggested, often informally, in the literature. The classes of models we consider cover three broad types of motivations: a direct cost of lying (e.g., Ellingsen and Johannesson 2004, Kartik 2009); a reputational cost derived from the belief that an audience holds about the subject's traits or action (e.g., Mazar et al. 2008), including guilt aversion (e.g., Charness and M. Dufwenberg 2006); and the influence of social norms and social comparisons (e.g., Weibull and Villa 2005). We also consider numerous extensions, combinations and mixtures of the aforementioned models (e.g., Kajackaite and Gneezy 2017). For all models we make minimal assumptions on the functional form and allow for heterogeneity of preference parameters, thus allowing us to derive very general conclusions.

Our empirical strategy to test the validity of the proposed explanations proceeds in two steps. First, we check whether each model is able to match the stylized findings of the meta study. This rules out many models, including models where the individual only cares about their reputation of having reported truthfully. In these models individuals are often predicted to pool on the same report, whereas the meta study shows that this is never the case. However, we also find eleven models that can match all the stylized findings of the meta study. These models offer very different mechanisms for the aversion to lying with very different policy implications. It is therefore important to be able to make sharper distinctions between the models. In the second step, we thus design four new experimental tests that allow us to further separate 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[5]; (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. Our predictions come in two varieties: (i) to (iii) are comparative statics while (iv) concerns properties of equilibrium behavior.

We take a Popperian approach in our empirical analysis (Popper 1934). Each of our

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.

[5] 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).

tests, taken in isolation, is not able to pin down a particular model. For example, among the models we consider, there are at least three very different motives that are consistent with the behavior we find in test (i), namely a reputation for honesty, inequality aversion and disappointment aversion. 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, 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 Appendices F and G), we implement our four tests in new laboratory experiments with more than 1600 subjects. To test the influence of the distribution of true states (test (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. 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.

To test the influence of subjects' beliefs about what others report (test (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" that 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 comparison 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 (test (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 material consequences of lying, e.g., being excluded from future experiments. We find significantly less over-reporting in the treatment in which the true state is observable compared to when it is not. This finding is again inconsistent with direct lying cost models and social comparison models since in those models utility does not depend on the observability of the true state. Moreover, we find that no subject lies downwards in this treatment (test (iv)).

In Section 4, we compare the predictions of the models to the gathered empirical evidence. The main empirical finding is that our four tests rule out almost all of the models previously suggested in the literature. Of the models we propose and consider, only two cannot be falsified by our data. Both models combine a preference for being seen as honest with a preference for being honest. This combination is also present in the concurrent papers by Khalmetski and Sliwka (forthcoming) and Gneezy et al. (2018). Both papers assume that individuals want to be perceived as honest and suffer from a lying cost related to the material gain from lying. A distinct intuition is explored in another concurrent paper by M. Dufwenberg and M. A. Dufwenberg (2018), who suppose that individuals care about the perception about by how much they have cheated, i.e., lied for material gain. We discuss how these studies relate to ours in the Conclusions. We then turn to calibrating a simple, linear version of one of our non-falsified models, showing that it can quantitatively reproduce the data from the meta study as well as the patterns in our new experiments. In the model, individuals suffer a fixed cost of 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 a preference for being seen as honest and a preference for being honest are the main motivations for truth-telling. 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 90 experimental studies containing 429 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 (i.e., we require at least two states) 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. We only consider distributions that either (i) have more than two states and are uniform or symmetric single-peaked, or (ii) have two states (with any distribution). This excludes only a handful of treatments in the literature. For more details on the selection process, see Appendix A.

We contacted the authors of the identified papers and obtained the raw data of 54 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 270616 decisions by 44390 subjects. Experiments were run in 47 countries which cover 69 percent of world population and 82 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 1 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. 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. Most importantly, we can conduct analyses that the original authors did not conduct. For other meta studies using the full individual subject data (albeit on different topics), see e.g., Harless and Camerer (1994), Weizsäcker (2010) or Engel (2011).

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 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:

$$\begin{array}{ll} r_{standardized} = \frac{\pi - E[\pi^{truthful}]}{E[\pi^{truthful}] - \pi^{min}} & \text{if} \quad \pi < E[\pi^{truthful}] \\ r_{standardized} = \frac{\pi - E[\pi^{truthful}]}{\pi^{max} - E[\pi^{truthful}]} & \text{if} \quad \pi \geq E[\pi^{truthful}] \end{array}$$

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 $E[\pi^{truthful}]$ 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.

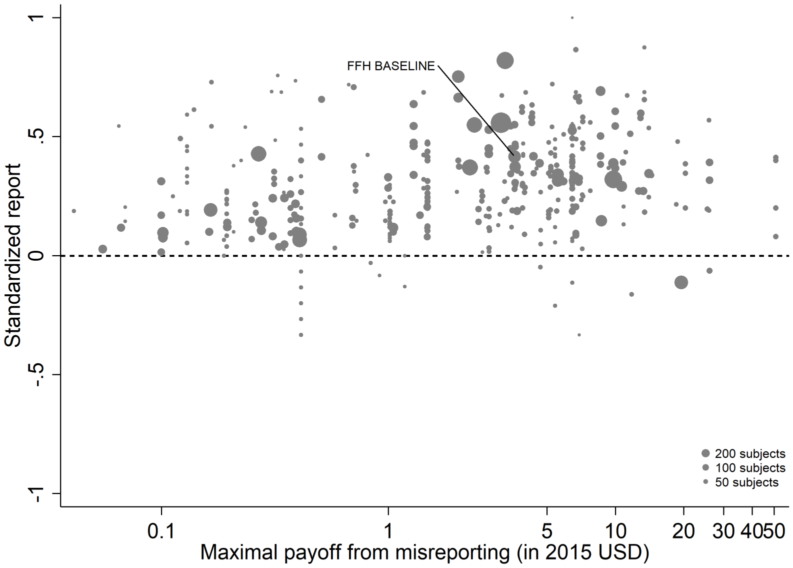
In general, without making further assumptions, one cannot say how many people lied or by how much in the FFH paradigm. We can only say how much money people left on the table. An average standardized report greater than 0 means that subjects leave less money on the table than a group of subjects who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

To give readers the possibility to explore the data in more detail, we have made interactive versions of all meta-study graphs available at www.preferencesfortruthtelling.com. The graphs allow restricting the data, e.g., only to specific countries. The graphs also provide more information about the underlying studies and give direct links from the plots to the original papers.

1.2 Results

Finding 1 *The average report is bounded away from the maximal report.*

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).

Figure 1 depicts an overview of the data. Standardized report is on the y-axis and the maximal payoff from misreporting, 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 monetary-payoff maximizers and had no concerns about lying, all bubbles would be at +1. In contrast, we find that the average standardized report is only 0.234. This is significantly ($p < 0.001$) lower than 0.25 or any higher threshold (clustering on subject; 0.38 when clustering on study) and thus bounded away from 1. This means that subjects forego about three-quarters of the potential gains from lying. This is a very strong departure from the standard economic prediction.

This finding turns out to be quite robust. Subjects continue to refrain from lying maximally when stakes are increased. 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 (Kajackaite and Gneezy 2017), a 500-fold increase. 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 -0.005 (using between-study variation as in Figure 1) or 0.003 (using within-study variation). See Appendix A for more details and for a comparison of our different identification strategies. This means that increasing incentives even further is unlikely to yield the standard economic prediction of +1. In Appendix A, we also show that subjects still refrain from lying maximally when they report repeatedly. In fact, repetition is associated with significantly lower reports. Learning and experience thus do not diminish the effect. Reporting behavior is also quite stable across countries and adding country fixed effects to our main regression (see Table A.2) increases the adjusted R^2 only from 0.370 to 0.457.

We next analyze the distribution of reports within each treatment.

Finding 2 *For each distribution of true states, more than one state is reported with positive probability.*

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 68 percent of all subjects in the meta study (the vast majority of the remaining subjects are in treatments with non-uniform distributions – where Finding 2 also holds). 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. The bold line is the average across all treatments, the grey area around it the 95% confidence interval of the average. As one can see in Figure 2, all possible reports are made with positive probability in almost all treatments. More generally, for each distribution of true states we have data on, the likelihood of the modal report is significantly ($p < 0.001$) lower than 0.79 (or any higher threshold), and thus bounded away from 1. We have enough data to cluster on study for the two distributions in Figure 2 and the result is robust to such clustering.

Finding 3 *When the distribution of true states is uniform, the probability of reporting a given state is weakly increasing in its payoff.*

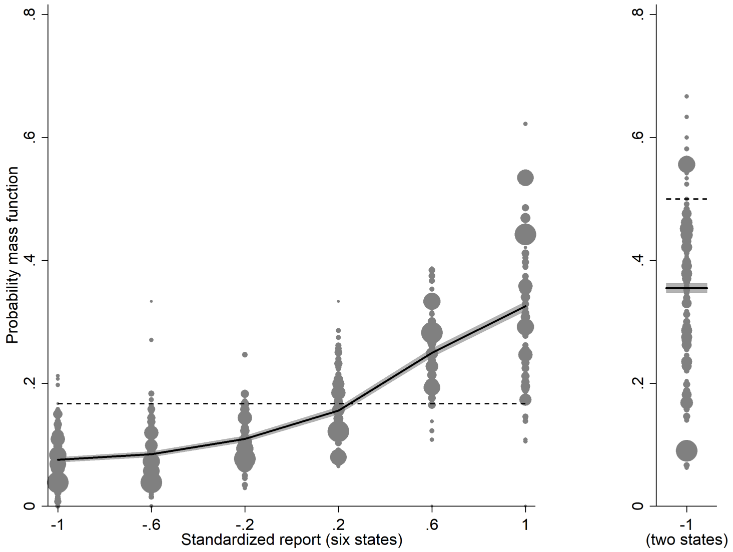
The figure also shows that reports that lead to higher payoffs are generally made more often, both for 6-state and 2-state distributions. The right panel of Figure 2 plots the likelihood of reporting the low-payoff state (standardized report of -1) for 2-state experiments. The vast majority of the bubbles are below 0.5 which implies that the high-payoff report is above 0.5. This positive correlation between the payoff of a given state and its likelihood of being reported holds for all uniform distributions we have data on (OLS regressions, all $p < 0.001$). We have enough data for the distributions with 2, 3, 6, and 10 states to test report-to-report changes and find that the reporting likelihood is strictly increasing for 2, 3 and 6 states (all $p < 0.008$) and weakly increasing for 10 states. We have enough data to cluster on study for 2- and 6-state distributions and the result is robust to such clustering.

Finding 4 *When the distribution of true states has more than 3 states, some non-maximal-payoff states are reported more often than their true likelihood.*

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 6-state experiments is more likely than 1/6 in almost all treatments. Such over-reporting of non-maximal states occurs in all distributions with more than three states we have data on (see Figure A.7 for the uniform distributions). We test all non-maximal states that are over-reported against their truthful likelihood using a binomial test. The lowest p-value is smaller than 0.001 for all distributions (we exclude distributions for which we have very little data, in particular, only one treatment). We have enough data to cluster on study for the uniform 6-state distribution and the result is robust to such clustering.

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. The bold line is the average across all treatments, the grey area around it the 95% confidence interval of the average.

Table 1: List of studies included in the meta study

Study	# treatments	# subjects	Country	Randomization method	True distribution
this study *	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 *	3	507	Germany	draw from urn	1D10
Amir et al. 2016 *	11	403	Israel	coin toss	20D2
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. 2017	2	120	Germany	coin toss	2D2
Banerjee et al. 2018 *	8	672	India	die roll	1D6
Barfort et al. 2015	1	862	Denmark	die roll	asy. 1D2
Basic et al. 2016 *	3	272	Germany	die roll	1D6
Beck et al. 2018 *	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
Cappelen et al. 2016 *	2	1473	Tanzania	coin toss	6D2
Charness et al. forthcoming	4	338	Spain	die roll	1D10
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 forthcoming	1	162	Switzerland	coin toss	1D2
Cohn et al. 2018	4	468	Switzerland	coin toss	1D2
Conrads et al. 2013 *	4	554	Germany	die roll	1D6
Conrads and Lotz 2015 *	4	246	Germany	coin toss	4D2
Conrads et al. 2017	1	114	Germany	die roll	1D2
Dai et al. 2018	2	384	France	die roll	1D3
Dato and Nieken 2016	1	288	Germany	die roll	1D6
Dieckmann et al. 2016	5	1015	multiple (5)	coin toss	1D2
Diekmann et al. 2015 *	1	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
Duch and Solaz 2016	3	3400	multiple (3)	die roll	1D6
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

Study	# treatments	# subjects	Country	Randomization method	True distribution
Gächter and J. F. Schulz 2016a *	23	2568	multiple (23)	die roll	1D6
Gächter and J. Schulz 2016 *	4	262	United Kingdom	die roll	1D6
Garbarino et al. forthcoming	3	978	USA	coin toss	multiple
Gino and Ariely 2012	8	304	USA	die roll	1D6
Gneezy et al. 2018	2	207	Germany	draw from urn	multiple
Grigorieff and Roth 2016 *	2	1511	USA	coin toss	4D2
Halevy et al. 2014 *	1	51	Netherlands	die roll	1D6
Hanna and Wang 2017	2	826	India	die roll	1D6
Heldring 2016 *	1	415	Rwanda	coin toss	30D2
Hilbig and Hessler 2013 *	6	765	Germany	die roll	asy. 1D2
Hilbig and Zettler 2015 *	4	342	Germany	multiple	asy. 1D2
Houser et al. 2012	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 2016 *	30	1390	multiple (15)	coin toss	1D2
Jacobsen and Piovesan 2016	3	148	Denmark	die roll	1D6
Jiang 2013 *	2	39	Netherlands	die roll	1D2
Jiang 2015 *	4	224	multiple (4)	die roll	1D2
Kajackaite and Gneezy 2017	17	1303	multiple (2)	multiple	multiple
Kroher and Wolbring 2015 *	7	384	Germany	die roll	1D6
Lowes et al. 2017	1	499	DR Congo	die roll	30D2
Maggian and Montinari 2017	2	192	France	die roll	1D2
Mann et al. 2016	10	2179	multiple (5)	die roll	1D2
Meub et al. 2016	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 *	2	102	Netherlands	draw from urn	1D2
Rauhut 2013 *	3	240	Switzerland	die roll	1D6
Ruffle and Tobol 2014 *	1	156	Israel	die roll	1D6
Ruffle and Tobol 2014 *	1	427	Israel	die roll	1D6
Schindler and Pfattheicher 2017 *	2	300	USA	coin toss	1D2
Shalvi et al. 2011 *	2	129	USA	die roll	1D6
Shalvi 2012	2	178	Netherlands	coin toss	20D2
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 *	4	120	Netherlands	coin toss	1D2
Shen et al. 2016	1	205	Singapore	die roll	1D6
Škoda 2013	3	90	Czech Republic	die roll	1D6
Suri et al. 2011	3	674	multiple (2)	die roll	multiple
Thielmann et al. 2017 *	1	152	Germany	coin toss	asy. 1D2
Utikal and Fischbacher 2013	2	31	Germany	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.

2 Theory

The meta study shows that subjects display strong aversion to lying and that this results in specific patterns of behavior as summarized by our four findings. In this section, we use a unified theoretical framework to formalize various ways that could potentially explain these patterns (introduced in Section 2.1). In order to address the breadth of plausible explanations and to be able to draw robust conclusions, we consider a large number of potential mechanisms, most of them already discussed, albeit often informally, in the literature. Indeed, one key contribution of our paper is to formalize in a parallel way a variety of suggested explanations. 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 (e.g., they care about the beliefs of some audience that observes their report); or they care about social comparisons or social norms which affect the reporting decision. In Section 2.2, we discuss one example model for each of the three types of explanations, including one of the two models that our empirical exercise will not be able to falsify. We discuss the remaining models in the appendices.

To test the models against each other, we first check whether they are able to explain the stylized findings of the meta study (Section 2.3). We find that many different models can do so. We therefore use our theoretical framework to develop four new tests that can distinguish between the models consistent with the meta study (Section 2.4). Table 2 lists all models and their predictions. For comparison purposes, we also state the results of our experiments in the row labeled Data.

2.1 Theoretical Framework

An individual observes state $\omega \in \Omega_n$, drawn i.i.d. across individuals from distribution F (with probability mass function f). We will suppose, except where noted, that the drawn state is observed privately by the individual. We suppose Ω_n is a subset of equally spaced natural numbers from ω_1 to ω_n, ordered $\omega_1, \omega_2, \ldots, \omega_n$ with $n > 1$. As in the meta study, we only consider distributions F that have $f(\omega) \in (0,1)$ for all $\omega \in \Omega_n$ and that either (i) have more than two states and are uniform or symmetric single-peaked, or (ii) have two states (with any distribution). Call this set of distributions $\mathcal{F}$.[6] After observing a state, individuals publicly give a report $r \in R_n$, where R_n is a subset of equally spaced natural numbers from r_1 to r_n, ordered $r_1, r_2, \ldots, r_n$. Individuals receive a monetary payment which is equal to their report r. We suppose that there is a natural mapping between each element of R_n and the corresponding element of Ω_n.[7] For example, imagine an individual privately flipping a coin. If they report heads, they receive £10, if they report tails they receive nothing. Then $\omega_1 = r_1 = 0$, and $\omega_2 = r_2 = 10$. We denote the distribution over reports as G (with probability mass function g). An individual is a liar if they report $r \neq \omega$. The proportion of liars at r is $\Lambda(r)$.

We denote a utility function as ϕ. 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 do maintain continuity of ϕ). We will also suppose, except where specifically noted, that sub-functionals of ϕ are continuous in their arguments.

We suppose that individuals are heterogeneous. They have a type $\vec{\theta} \in \Theta$, where $\vec{\theta}$ is a vector with J entries, and Θ is the set of potential types $\times_J [0, \kappa^j]$, with $\kappa^j \in \mathbb{R}^{++}$. Each of the J elements of $\vec{\theta}$ gives the relative trade-off experienced by an individual between monetary benefits and specific non-monetary, psychological costs (e.g., the cost of lying, or reputational costs). When we introduce specific models, we will only focus on the subvector of $\vec{\theta}$ that is relevant for each model (which will usually contain only one or two entries). We suppose that $\vec{\theta}$ is drawn i.i.d. from H, a non-atomic distribution on Θ. Each entry θ^j is thus distributed on $[0, \kappa^j]$.[8] In Appendix E, we show that the set of non-falsified models does not change if we assume that H is degenerate. The exogenous elements of the models are thus the distribution F over states and the distribution H over types while the distribution G over reports and thus the share of liars at r, $\Lambda(r)$, arise endogenously in equilibrium.

[6] A handful of papers in the meta study use non-equally spaced states. All our results also hold for these distributions and for any distribution where the payoffs are not "too" unequally spaced.

[7] Formally, we can think of there being as an order-preserving bijection between Ω_n and R_n. A simpler (albeit slightly less general) conceptualization is that a report is the identity function from Ω_n to itself.

[8] Our assumptions on κ^j and H imply that our framework for more general models does not nest, strictly speaking, the standard model, where individuals only care about their monetary payoff. Instead, the standard model is a limit case of our models (where the κs go to 0, or the support of H becomes concentrated on 0). This allows the predictions generated by more general models to be sharply distinguished from the predictions of the standard model (as opposed to nesting them). The same reasoning applies to other "nested" models, e.g. the lying cost (LC) model is a limit case of the Reputation for Honesty + LC model.

We assume that individuals only report once and there are no repeated interactions. We suppose a continuum of "subject" players and a single "audience" player (the continuum of subjects ensures that any given subject has a negligible impact on the aggregate reporting distribution). The subjects are individuals exactly as described above. The audience takes no action, but rather serves as a player who may hold beliefs about any of the subjects after observing the subjects' reports. The audience could, e.g., be the experimenter or another person the subject reveals their report to. Subjects do not observe each others' reports. Utility may depend on the distribution of others' reports, the drawn state-report combinations of others, or beliefs.[9] Because subjects take a single action we can consider a strategy as mapping type and state combinations ($\vec{\theta} \times \omega$) into a distribution over reports r.[10] When an individual's utility depends on the beliefs of other players, we consider the Sequential Equilibria of the induced psychological game, as introduced by Battigalli and M. Dufwenberg (2009). (The original psychological game theory framework of Geanakoplos et al. (1989) cannot allow for utility to depend on updated beliefs.). When utility does not depend on others' beliefs, the analysis can be simplified and we assume the solution concept to be the set of standard Bayes Nash Equilibria of the game. In some of our models, an individual's utility depends only on their own state and report. In this case, our solution concept is simply individual optimization, but for consistency, we also use the words equilibrium and strategy to describe the outcomes of these models.

2.2 Modelling Preferences for Truth-telling

In this section, we introduce one example for each of the three main categories of lying aversion: lying costs (Section 2.2.1), social norms/comparisons (2.2.2), and reputational concerns (2.2.3). The remaining models are described in Appendix B. Some of these models represent other ways of formalizing the effect of descriptive norms and social comparisons on reporting,

[9] Our approach is similar to population games in many ways; for example, in that we have a continuum of agents (see Sandholm (2015) for a summary of population games). However, in many models utility may depend not just on the aggregate distribution of reports, but also the relationship between a given report and its associated drawn state.

[10] 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 $\vec{\theta}$ and the non-monetary costs 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. Because subjects in the experiment are anonymous to each other we also only focus on equilibria where strategies cannot depend on the identity of the player (but of course, it can depend on their preference parameters).

including a model of inequality aversion (Appendix B.1); a model that combines lying costs with inequality aversion (B.2); and a social comparisons model in which only subjects who could have lied upwards matter for social comparisons (B.3). Other models build on the idea of reputational concerns and include a model where individuals want to signal to the audience that they place low value on money (B.4); a model where individuals want to cultivate a reputation as a person who has high lying costs (B.5); and a model of guilt aversion (B.6). Finally, we include a model of money maximizing with errors (B.7), and a model that combines lying costs with expectations-based reference-dependence (B.8). In addition, Appendix C describes several models that fail to explain the findings of the meta study and that are therefore not further considered in the body of the paper. Most prominently, we discuss a model in which individuals only care about the audience's belief about their honesty (Appendix C.2).

2.2.1 Lying Costs (LC)

A common explanation for the reluctance to lie is that deviating from telling the truth is intrinsically 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)[11]; from "injunctive" social norms of honesty, i.e. norms that are based on a shared perception that lying is socially disapproved; or from the unwillingness to defy the authority of the person or institution who asks for the private information. Such "lying cost" (LC) models have wide popularity in applications and represent a simple extension of the standard model in which individuals only care about their monetary payoff. Our formulation of this class of models 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.[12]

[11] 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 Appendix C, 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.

[12] 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); and DellaVigna et al. (2016).

Formally, we suppose individuals have a utility function

$$\phi(r, c(r, \omega); \theta^{LC})$$

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 $r = \omega$ which is not necessarily unique. (For some specifications, for example fixed costs of lying, c will not be differentiable in its arguments.) For our calibrational exercises we normalize $c(\omega, \omega) = 0$, so that individuals experience no cost when they tell the truth. In order to make the model non-trivial we suppose that there is at least one non-maximal state ω such that there exists an $r > \omega$ where $c(r, \omega) > c(\omega, \omega)$ (otherwise no one would ever pay any costs to lying). The only element of $\vec{\theta}$ that affects utility is the scalar θ^{LC} which governs the weight that an individual applies to the lying cost. We make a few assumptions on ϕ. First, ϕ is strictly increasing in the first argument, fixing all the other arguments; this captures the property that utility is increasing in the monetary payment received. Second, ϕ is decreasing in the second argument, fixing all the other arguments, capturing the property that utility falls as the cost of lying increases. In particular, it is strictly decreasing for all $\theta^{LC} > 0$. Third and fourth, fixing all other arguments, ϕ is (weakly) decreasing in θ^{LC}, and the cross partial of ϕ with respect to c and θ^{LC} is strictly negative, while other cross partials are 0. This captures the properties that an individual with a higher draw of θ^{LC} 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^{LC}.[13] The solution to LC models can be found by simply solving a single decision maker's optimization problem.

2.2.2 Social Norms: Conformity in LC

Another potential explanation for lying aversion extends the intuition of the LC model. It posits that individuals care about social norms or social comparisons which inform their

[13] Our results regarding the LC model can be easily generalized further: they do not require that utility is weakly decreasing in θ^{LC}, only that the restriction on the cross partials hold. We make the assumption that utility is weakly decreasing in θ^{LC} as it allows for a natural interpretation of θ^{LC} (the same applies to the following models). Our results also do not depend on individuals all having the same functional form c so long as the assumptions regarding θ^{LC} hold. So, for example, our results hold when some individuals have fixed and others convex costs of lying.

reporting decision. The leading example is that individuals may feel less bad about lying if they believe that others are lying too. Importantly, the norms here are "descriptive" in the sense that they are based on the perception of what others normally do, rather than "injunctive", which are instead based on the perception of what ought to be done and do not depend on the behavior of others (injunctive norms are better captured by LC models). We call such a model "Conformity in LC". 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). We suppose that an individual's total utility loss from misreporting depends both on an LC cost (as described in the previous model), but also on the average LC cost in society. The latter depends not just on players' actions, but the profile of joint state-report combinations across all individuals. Because we can think of any individuals' drawn state as part of their privately observed type, we use the framework of Bayes Nash Equilibrium.[14]

Formally, in the Conformity in LC model individuals have a utility function

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

$c(r,\omega)$ has the same interpretation and assumptions as in the LC model and types are heterogeneous in the scalar θ^{CLC} (where CLC denotes the "Conformity in LC" model specific parameter; analogous abbreviations are used for the rest of the models); the rest of the vector $\vec{\theta}$ again does not affect utility. $\bar{c}$ is the average incurred LC cost in society. This average cost is determined in equilibrium, and thus all individuals know what it is; for notational ease we supress the dependence of $\bar{c}$ on the other parameters of the model. η captures the "normalized cost of lying", i.e., the cost of lying conditional on the incurred LC cost in society (for our calibrational exercises we suppose $\eta(0,\bar{c}) = 0$) and is strictly increasing in its first argument. For $c > 0$, η is strictly falling in the second argument so that the normalized cost is increasing in the individual's own personal lying cost and falling in the aggregate LC cost, i.e., their lying costs are falling as others lie more (for $c = 0$, the partial of η with respect to its second

[14] Since we suppose a continuum of agents, one can also think of utility as depending on the strategies of others (integrating out over θ^{CLC}). Observe that we suppose in this model that individuals' utility depends on the actual costs of others. An alternative framing would be where the utility for an individual depends on their own beliefs about others' costs. With a continuum of agents, and correct beliefs, these equal the realized costs.

argument is 0). As in the previous model ϕ is strictly increasing in its first argument, and decreasing in the second argument (strictly so for all $\theta^{CLC} > 0$). ϕ is (weakly) decreasing in θ^{CLC} fixing the first two arguments, and the cross partial of ϕ with respect to η and θ^{CLC} is strictly negative, while other cross partials are 0. These assumptions are analogous to the ones presented in the previous models and capture the same intuitions.

2.2.3 Reputation for Honesty + LC

A different way to extend the LC model is to allow individuals to experience both an intrinsic cost of lying, as well as reputational costs associated with inference about their honesty (e.g., Khalmetski and Sliwka forthcoming, Gneezy et al. 2018). We suppose that an individual's utility is falling in the belief of the audience player that the individual's report is not honest, i.e., has a state not equal to the report. Akerlof (1983) provides the first discussion in the economics literature that honesty may be generated by reputational concerns and many recent papers have built on this intuition.[15] Thus, an individual's utility is belief-dependent, specifically depending on the audience player's updated beliefs. Thus, we must use the tools of psychological game theory to analyze the game. We use the framework of Battigalli and M. Dufwenberg (2009) in our analysis.[16] Of course, the audience cannot directly observe whether a player is lying, and has to base their beliefs on the observable report r. Utility is thus a decreasing function of the audience's belief about whether an individual lied. Because the audience player makes correct Bayesian inference based on observing the report and knowing the equilibrium strategies, their posterior belief about whether an individual is a liar, conditional on a report r, is $\Lambda(r)$, the fraction of liars at r in equilibrium. We therefore directly assume that utility depends on $\upsilon(\Lambda(r))$, with υ a strictly increasing function.

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

[15] This includes, for example, Mazar et al. (2008); Suri et al. (2011); Hao and Houser (2017); Shalvi and Leiser (2013); Utikal and Fischbacher (2013); Fischbacher and Föllmi-Heusi (2013); Gill et al. (2013) and Hilbig and Hessler (2013).

[16] Some researchers have suggested that a simple model in which individuals care only about the audience's belief that they are a liar, conditional on their report, could explain behavior. We discuss in Appendix C.2 why such a model fails to match the findings of the meta study, and why reputational concerns need to be combined with some other motive to explain the data within our theoretical framework. A related model by M. Dufwenberg and M. A. Dufwenberg (2018) posits that individuals care about the inferred degree of over-reporting. This model builds on different distributional assumptions than those we use in our paper. We discuss the role of distributional assumptions for our results in Appendix E.

image (i.e., the audience's beliefs). We focus on a situation where there is additive separability between the different components of the utility function.[17] Formally, in the "Reputation for Honesty + LC" model utility is

$$\phi(r, c(r,\omega), \Lambda(r); \theta^{LC}, \theta^{RH}) = u(r) - \theta^{LC} c(r,\omega) - \theta^{RH} \upsilon(\Lambda(r))$$

u is strictly increasing in r. Types are heterogeneous in the scalars θ^{LC} and θ^{RH} and the rest of $\vec{\theta}$ does not affect utility. c is as described in the LC model. υ is a strictly increasing function of $\Lambda(r)$ with a minimum at 0 (and in calibrational exercises we normalize $\upsilon(0) = 0$). Thus the individual likes more money, but dislikes lying and being perceived as a liar by the audience. The functional form implies analog patterns for the cross partials as the previous models.[18]

2.3 Distinguishing Models Using the Meta Study

We now turn to understanding how our models can be distinguished in the data. The first test is whether the models can match the four findings of the meta study. We find that the three models presented in the previous section, as well as all those listed in Appendix B, can do so.

Proposition 1 *There exists a parameterization of the LC model, the Conformity in LC model, the Reputation for Honesty + LC model and of all other models listed in Appendix B (i.e., Inequality Aversion; Inequality Aversion + LC; Censored Conformity in LC; Reputation for Being Not Greedy; LC-Reputation; Guilt Aversion; Choice Error; and Kőszegi and Rabin + LC) which can explain Findings 1–4 for any number of states n and for any $F \in \mathcal{F}$.*

All proofs for the results in this section are collected in Appendix D. The proof for the LC model constructs one example utility function, combining a fixed cost and a convex cost of lying, and then shows that it yields Findings 1–4 for any n and any $F \in \mathcal{F}$. Many of the other models considered in this paper contain the LC model as limit case and can therefore

[17] A similar additive-separability assumption has been used in related papers combining intrinsic lying costs and reputational concerns (Khalmetski and Sliwka forthcoming; Gneezy et al. 2018).

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

explain Findings 1–4. However, there are several models, e.g., the Inequality Aversion model (Appendix B.1) or the Reputation for Being Not Greedy model (B.4), which rely on very different mechanisms and can still explain Findings 1–4.

2.4 Distinguishing Models Using New Empirical Tests

Proposition 1 shows that the existing literature, reflected in the meta study, cannot pin down the mechanism which generates lying aversion. The meta study does falsify quite a few popular models, which we discuss in Appendix C, but the data is not strong enough to narrow the set of surviving models further down. This motivates us to devise four additional empirical tests which can distinguish between the models that are in line with the meta study. Three of the four new tests are "comparative statics" and one is an equilibrium property: (i) how does the distribution of true states affect the distribution of reports; (ii) how does the belief about the reports of other subjects influence the distribution of reports; (iii) does the observability of the true state affect the distribution of reports; (iv) will some subjects lie downwards if the true state is observable. 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.[19]

We derive predictions for each model and for each test using very general specifications of individual heterogeneity and the functional form. We present predictions for an arbitrary number of states n and for the special case of $n = 2$. On the one hand, allowing for an arbitrary number of states generates predictions that are applicable to a larger set of potential settings. On the other hand, restricting $n = 2$ allows us to make sharper predictions, and thus potentially falsify a larger set of models. For example, for models where individuals care about what others do (e.g., social comparison models), it doesn't matter whether individuals care about the average report or the distribution of reports when $n = 2$. For models that rule out downward lying, the binary setting also 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 number of

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

Table 2: Summary of Testable Predictions

Model		**New Tests**				**Section**
	Can Explain Meta Study	**Shift in True Distribution F**	**Shift in Belief About Reports $\hat{G}$**	**Observability of True State ω**	**Lying Down Unobs./Obs.**	
Lying Costs (LC)	Yes	f-invariance	$\hat{g}$-invariance	o-invariance	No/No	2.2.1
Social Norms/Comparisons						
Conformity in LC*	Yes	drawing out	affinity	o-invariance	No/No	2.2.2
Inequality Aversion*	Yes	f-invariance	affinity	o-invariance	Yes/Yes	B.1
Inequality Aversion + LC*	Yes	drawing in	affinity	o-invariance	-/-	B.2
Censored Conformity in LC*	Yes	f-invariance	affinity	o-invariance	No/No	B.3
Reputation						
Reputation for Honesty + LC*	Yes	drawing in	-	o-shift	-/No	2.2.3
Reputation for Being Not Greedy*	Yes	f-invariance	-	o-invariance	Yes/Yes	B.4
LC-Reputation*	Yes	drawing in	-	o-shift	-/-	B.5
Guilt Aversion*	Yes	f-invariance	affinity	o-invariance	Yes/Yes	B.6
Choice Error	Yes	f-invariance	$\hat{g}$-invariance	o-invariance	Yes/Yes	B.7
Kőszegi-Rabin + LC	Yes	-	$\hat{g}$-invariance	o-invariance	No/No	B.8
Data		**drawing in**	**$\hat{g}$-invariance**	**o-shift**	**?/No**	

Notes: The details of the predictions are explained in the text. "-" means that, depending on parameters, any behavior can be explained. The predictions for shifts in F and $\hat{G}$ are for 2-state distributions, i.e., $n = 2$. 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.

observed high-payoff reports and we are left with the reports made by the subjects who have drawn the low-payoff state. Moreover, conducting our new tests with 2-state distributions is simpler and easier to understand for subjects. Recall that across all results, we only consider distributions $F \in \mathcal{F}$.

The models, as well as the predictions they generate in each of the tests, are listed in Table 2. We report the 2-state predictions in the columns describing the effect of shifts in the distributions of true states F and beliefs about others' reports $\hat{G}$ (see below for details), since we use 2-state distributions in our new experimental tests of these predictions. Some of the models we consider do not guarantee a unique reporting distribution G without additional parametric restrictions. We discuss below 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 2. 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 of equilibria.

We now turn to discussing our four empirical tests. The first test is about how the distribution of reports G (recall that $g(r_j)$ gives the unconditional fraction of individuals giving report r_j) changes when the higher states are more likely to be drawn (but while maintaining the same set of support for the distribution). Specifically we suppose that we induce a shift in the distribution of states F (recall that $f(\omega_j)$ gives the probability that state ω_j is drawn) that satisfies first order stochastic dominance. We then look at one minus the ratio of the observed number of reports of the lowest state to the expected number of draws of the lowest state: $\frac{f(\omega_1)-g(r_1)}{f(\omega_1)} = 1 - \frac{g(r_1)}{f(\omega_1)}$. For those models in which no individual lies downwards we can interpret the statistic as the proportion of people who draw ω_1 but report something higher, i.e., $r > r_1$.

Definition 1 *Consider two pairs of distributions: F^A, G^A and F^B, G^B where G^j is the reporting distribution associated with F^j, and where F^B strictly first order stochastically dominates F^A and they all have full support. A model exhibits* drawing in/drawing out/f-invariance *if $1 - \frac{g^B(r_1)}{f^B(\omega_1)}$ is larger than/smaller than/the same as $1 - \frac{g^A(r_1)}{f^A(\omega_1)}$.*

Thus, the term "drawing in" means that the lowest state is even more underreported when higher states become more likely. “Drawing out” refers to the opposite tendency. As we will show below, several very different motivations can lead to drawing in. For example, increasing the true probability of high states increases the likelihood that a high report is true, leading subjects who care about being perceived as honest, as in our Reputation for Honesty + LC model (Section 2.2.3), to make such reports more often. But increasing the true probability of high states also increases the likelihood that other subjects report high, pushing subjects who dislike inequality (Appendix B.2) to report high states. And subjects who compare their outcome to their recent expectations (Appendix B.8) could also react in this way.[20]

The second test looks at how an individual's probability of reporting the highest state will change when we exogenously shift their belief about the distribution of reports. We will refer to $\hat{G}$ as the beliefs of players about the distribution of reports. In equilibrium, given correct beliefs about others, $\hat{G} = G$. Our experiment focuses on experimentally manipulating the beliefs about others, i.e. $\hat{G}$, so that they may no longer be correct, and then observing the resulting actual reporting distribution G. We focus on situations where there is full support on all reports in both beliefs and actuality.

Definition 2 *Fix a distribution over states F and consider two pairs of distributions $\hat{G}^A, G^A$ and $\hat{G}^B, G^B$, where G^j is the reporting distribution induced by F and by the belief that others will report according to $\hat{G}^j$. Moreover, suppose all exhibit full support and $\hat{G}^B$ strictly first order stochastically dominates $\hat{G}^A$. A model exhibits* affinity/aversion/$\hat{g}$-invariance *if $g^B(r_n)$ is larger than/smaller than/the same as $g^A(r_n)$.*

Thus, the term “affinity” means that reporting of the highest state increases when the subject believes that higher states are more likely to be reported by others. The term “aversion” refers to the opposite tendency. Such an exercise allows us to test the models in one of three ways. First, in some models, e.g., Inequality Aversion (Appendix B.1), individuals care directly about the reports made by others and thus $\hat{G}$ (or a sufficient statistic for it) directly enters the utility. Therefore, we can immediately assess the effect of a shift in $\hat{G}$ on

[20] 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 2 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 2 are based on the assumption of a unique equilibrium.

behavior.[21] For these models, shifting an individual's belief about $\hat{G}$ 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). These models all predict affinity.

Second, in some other models (Conformity in LC and Censored Conformity in LC), individuals care about the profile of joint state-report combinations across other individuals (i.e. the amount of lying by others). In these models no individual lies downwards and so, for binary states, $\hat{G}$ contains sufficient information about the joint state-report combinations. Thus, shifting $\hat{G}$ directly alters an individual's best response. These models again predict affinity.

Finally, this exercise allows us, albeit indirectly, to understand what happens when beliefs about H (the distribution of $\vec{\theta}$) change. Directly changing this belief is difficult since this requires identifying $\vec{\theta}$ for each subject and then conveying this insight to all subjects. However, for models with a unique equilibrium, because G is an endogenous equilibrium outcome, shifts in $\hat{G}$ can 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).[22] For many of these models, the conditions defining the unique equilibrium reporting strategy are invariant to shifts in $\hat{G}$ and H, which means that our treatment should not affect behavior. For another set of models, in particular Reputation for Being Not Greedy, Reputation for Honesty + LC and LC-Reputation, there is no simple mapping from $\hat{G}$ to beliefs about H and a shift in $\hat{G}$ could lead to affinity, aversion or $\hat{g}$-invariance.

Our third test considers whether or not it matters for the distribution of reports that the audience player can observe the true state. 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 *A model exhibits o-*shift *if G changes when the true state becomes observable to the audience, and o-*invariance *if G is not affected by the observability of the state.*

[21] Not all models can rationalize all Gs for a given F. We do not directly test whether subjects' predicted beliefs about distributions are allowed by any given model, given that we only elicit an average prediction of beliefs about reports.

[22] To specify the updating process more precisely, we suppose that individuals have a single probability distribution H 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 our results will continue to hold.

In some of the models we consider, the cost associated with lying are internal and therefore do not depend on whether an audience is able to observe the state or not. In other models, however, the costs depend on the inference the audience is able to make, and so observability of the true state affects predictions.[23]

Our fourth test comes in two parts. Both parts try to understand whether or not there are individuals who engage in downward lying, i.e., draw ω_i and report r_j with $j < i$. The first is whether downward lying can occur in an equilibrium with observability of the state by the audience and where G features full support. The second is an analogous test but in the situation where the state is not observed by the audience. We will only focus on the former test in our experiments.

Definition 4 *Fix a distribution over states F and an associated full-support distribution G over reports. The model exhibits downward lying if there exists some individual who draws ω_i but reports r_j where $j < i$. The model does not exhibit downward lying if there is no such individual.*

Although lying down may seem counter-intuitive, as we will show below, there can be a number of reasons why individuals may want to lie downwards. In models where individuals are concerned with reputation, lying downwards may be beneficial if low reports are associated with a better reputation than high reports. Alternatively, in models of social comparisons, such as the inequality aversion models, downward lying may arise because individuals aim to conform to others' reports.

The following proposition summarizes the predictions for the three models described above.

Proposition 2

- *Suppose individuals have LC utility. For an arbitrary number of states n, we have f-invariance, $\hat{g}$-invariance, o-invariance and no lying down when the state is unobserved or observed.*

- *Suppose individuals have Conformity in LC utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we may have affinity,*

[23] 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, just that the set of equilibria has changed.

aversion or $\hat{g}$-invariance, we have o-invariance and no lying down when the state is unobserved or observed. For $n = 2$, we have drawing out when the equilibrium is unique and we have affinity.

- *Suppose individuals have Reputation for Honesty + LC utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we may have affinity, aversion or $\hat{g}$-invariance, we have o-shift, depending on parameters, we may have lying down or not when the state is unobserved, and we have no lying down when the state is observed. For $n = 2$, we have drawing in when the equilibrium is unique.*

"Depending on parameters" refers to the distribution over states F, the distribution H over types, any sub-functions that might be introduced in a model definition, e.g., the cost function c in the LC model, and when considering affinity, aversion and $\hat{g}$-invariance, $\hat{G}$ (as this is something we experimentally manipulate). In the cases when predictions depend on parameters, the proofs will provide examples for each possible behavior. If the statement is unqualified, it means that it holds for any $F \in \mathcal{F}$, any H, sub-functions, and $\hat{G}$.

Before moving on, we provide some intuition for the results. For simplicity, we focus on 2-state/report distributions. In the LC model, individuals never lie downwards, because they (weakly) pay a lying cost and also receive a lower monetary payoff when doing so. Since only their own state and their own report matter for utility, conditional on drawing the low state, for a fixed $\vec{\theta}$, an individual will always make the same report, regardless of F or $\hat{G}$. Thus, we observe both f-invariance and $\hat{g}$-invariance. 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 is observed.

In the Conformity in LC models, individuals will never lie downwards since, as in the LC model, they would face a lower monetary payoff as well as a weakly higher cost of lying. Morever, with a unique equilibrium, as $f(\omega_2)$ increases, more individuals draw the high state and can report r_2 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_1, and was indifferent before between r_1 and r_2 will now strictly prefer r_1. This implies drawing out. In the Conformity in LC model, because G enters directly into the utility function and because no one lies downwards, we can tell how the individual's best response changes with shifts in expected G, i.e. $\hat{G}$. Fixing F, if $\hat{g}(r_2)$ 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, i.e., we have 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 is observed.

In the Reputation for Honesty + LC model, because individuals have a concern for reputation and also have lying costs, they may or may not lie down if the state is unobserved. If an individual is motivated relatively more by reputational concerns, then they will lie down if the state is unobserved. In contrast, if lying costs dominate as a motivation, they will not lie down. If the state is observed, no one lies downwards. Although multiple equilibria may occur, whenever the equilibrium is unique, the Reputation for Honesty + LC model exhibits drawing in. As $f(\omega_2)$ increases, some individuals who previously drew ω_1 will now draw ω_2. Those individuals now face a lower LC cost when giving the high report (which is in fact zero). Fixing the reputational cost, this implies some of them will now give the high report (instead of the low report). Fixing the behavior of others, this reduces the fraction of liars giving the high report and thus the reputational cost of the high report decreases; and similarly, increases the fraction of liars giving the low report. This reduces the (relative) cost of giving the high report even more. Therefore, we observe drawing in. Our intuition here relies on partial equilibrium reasoning, but the formal proof shows how to extend this to full equilibrium reasoning. Even with a unique equilibrium, we may observe either aversion, affinity or $\hat{g}$-invariance since it depends on how the distribution of H is perceived to have changed when $\hat{G}$ shifts.[24] Because the model includes reputational costs, whether or not the audience observes just the report, or also the state, matters for behavior.

In Appendix F, we provide additional evidence regarding predictions of the Kőszegi-Rabin

[24] If, for example, the change is interpreted as a shift by individuals who have low reputational costs, and so care mostly about LC costs, then an increase in $\hat{g}(r_2)$ will be interpreted as more individuals who drew ω_1 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 LC 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 LC costs are unlikely to ever give the high report). An increase in $\hat{g}(r_2)$ is then interpreted as individuals who drew ω_2 as being more willing to pay the reputation cost of reporting r_2. Thus, the fraction of truth-tellers at r_2 increases, driving affinity.

+ LC model which are not listed in the table. We also test specific f-invariance predictions for the LC model in a 10-state experiment, where we show that drawing-in like behavior also obtains in an experiment with 10 states.

3 New Experiments

In this section we report a large-scale (N = 1610) set of experiments designed to implement the four tests outlined above. The experiments were conducted with students at the University of Nottingham and University of Oxford. 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 G.

3.1 Shifting the Distribution of True States F

We test the effect of a shift in the distribution of true states F using treatments with 2-state distributions. 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 that serve to increase the length of the questionnaire so that the task appears meaningful. 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_1 and drawing a chip with 10 represents drawing ω_2. Reports of 4 and 10 are similarly r_1 and r_2. The chips are arranged on a tray on the subject's desk such that subjects are fully aware of the distribution F (see Appendix G 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.

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 tray 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). Note that the distribution used in F_HIGH first-order stochastically dominates the distribution in F_LOW in line with Definition 1. We select samples sizes such 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.

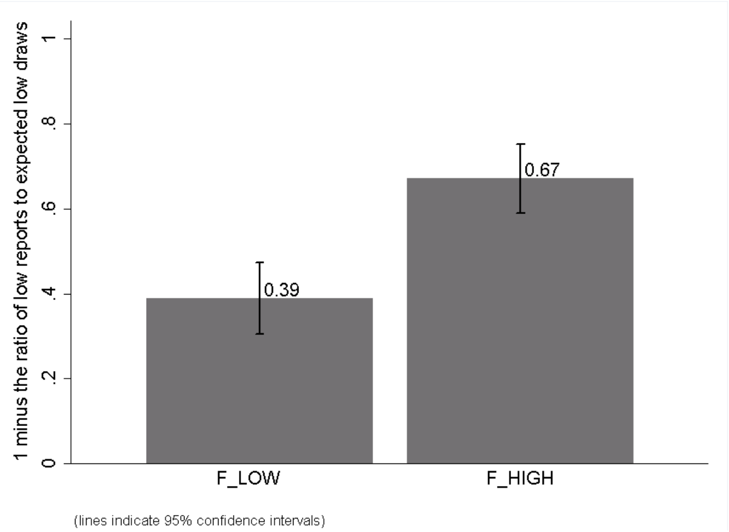
3.2 Results

Finding 5 *We observe drawing in, i.e., the statistic* $1 - \frac{g(r_1)}{f(\omega_1)}$ *is significantly higher in F_HIGH than F_LOW.*

Figure 3 shows the values of the statistic $1 - \frac{g(r_1)}{f(\omega_1)}$ 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 (this means that 45 percent of subjects in F_LOW and 87 percent in F_HIGH report 10). This difference of almost 30 percentage points is very large and highly significant ($p < 0.001$, OLS with robust SE; $p < 0.001$, χ^2 test).[25]

[25] This result is based on a pooled sample using observations collected both in Nottingham and Oxford. We obtain similar results if we focus on each sub-sample separately. Using only the Nottingham sub-sample (n = 391), we find a treatment difference of 28 percentage points ($p < 0.001$, OLS with robust SE; $p < 0.001$, χ^2 test). Using only the Oxford sub-sample (n = 83), we find a treatment difference of 27 percentage points ($p = 0.064$, OLS with robust SE; $p = 0.062$, χ^2 test).

Figure 3: Effect of shifting the distribution of true states



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., $\hat{G}$. There are three other studies testing the effect of beliefs on reporting (Rauhut 2013, Diekmann et al. 2015 and Gächter and J. Schulz 2016). These studies affect beliefs by showing to subjects the actual past behavior of participants. Diekmann et al. (2015) and Gächter and J. Schulz (2016) find no effect and Rauhut (2013) finds a positive effect. Rauhut (2013), however, compares subjects who have initially too high beliefs that are then updated downwards to subjects who have initially too low beliefs that are updated upwards. The treatment is thus not assigned fully randomly.

We use an alternative and complementary method. 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. The advantage of our design is that we do not need to sample selectively from the distribution of actual past behavior of other subjects. This could be problematic because, 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. 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.[26] 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 if their estimate is correct (within an error margin of ±3 percentage points). This mechanism is very simple and easier to explain and understand than proper scoring rules. It elicits in an incentive-compatible way the mode (or more precisely, the mid-point of the 6-percentage point interval with the highest likelihood) of a subject's distribution of estimates. We use subjects' estimates to check whether our anchoring manipulation is successful in shifting subjects' beliefs.[27]

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

[26] 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". Note that we did not ask subjects to guess or interpret the purpose of the experiment, but rather to reflect on participants' motives and satisfaction with various hypothetical behaviors undertaken in the experiment.

[27] For many distributions mode and mean are actually tightly linked. To illustrate this point, we have run the following simulation assuming that beliefs are distributed according to the very flexible beta distribution. We have generated 100000 pairs of beta distributions with randomly drawn α and β and compared the modes and means of the two distributions in each pair. In over 97 percent of cases where a mode exists and where one distribution has a higher mode than the other one, the higher-mode distribution has also a higher mean. This means that if our elicitation of the belief mode finds a difference between treatments, then it is highly likely that the two treatments also have different belief means.

May 2016 with a total of 340 subjects (173 in G_LOW, 167 in G_HIGH).

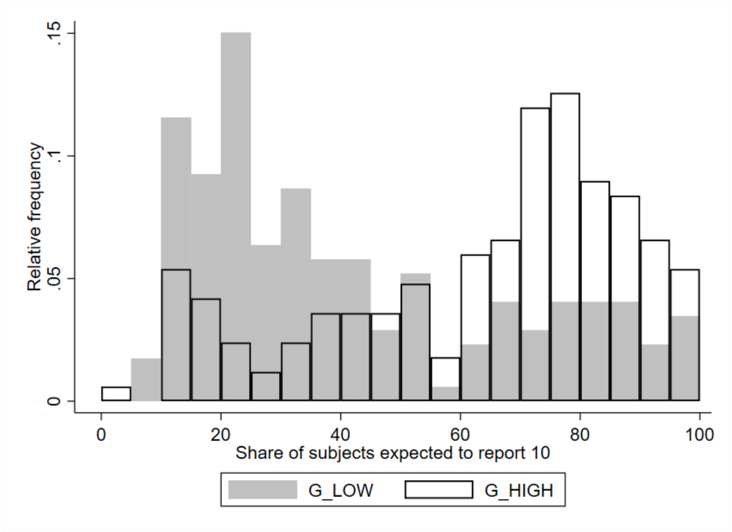
3.4 Results

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

Finding 6 *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 4 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, and practically first-order stochastically dominates it, in line with Definition 2. 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; $p < 0.001$, Wilcoxon rank-sum test).

Figure 4: 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.

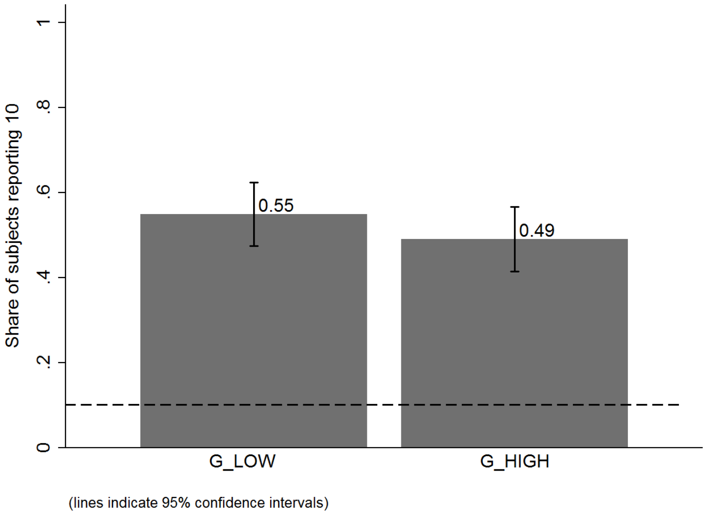
Finding 7 *The fraction of subjects reporting a 10 is not significantly different between G_HIGH* *and G_LOW, i.e., we cannot reject the null hypothesis of $\hat{g}$-invariance. The point estimate is in the direction of aversion.*

Figure 5 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; $p = 0.284$, χ^2 test). Taken together, our study and the previous literature provide converging evidence that manipulating beliefs about others' reports has a limited impact on reporting.

One word of caution is warranted. Even though the point estimate of the effect of the $\hat{G}$ treatments is quite close to zero, we cannot reject (small) positive or negative effects of a change in beliefs. A power analysis shows that we can only detect treatment differences of 15 percentage points or larger at the 5% level and with 80% power, but we are not sufficiently powered to detect small differences like that observed in Figure 5. 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, these models typically predict quite large responses to shifts in $\hat{G}$. For example, a simple, calibrated version of the Conformity in LC model implies that 21 percent of subjects should increase their reports across our $\hat{G}$ treatments, which we do have power to detect. In fact, our data show that (in net) 6 percent of subjects *decrease* their report.[28]

[28] 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 Conformity in LC 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 $\hat{G}$ treatments. For details of this calibration see Appendix H.1.

Figure 5: Effect of shifting beliefs about the distribution of reports



3.5 Changing the Observability of States

A final set of treatments tests whether observability of the subject's true state by the experimenter affects reporting behavior, in line with Definition 3. The experiments use a setup similar to the one 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. Differently from the previous experiments, 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).[29] Second, we introduce a payment procedure that

[29] 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

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.[30] We ran these 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

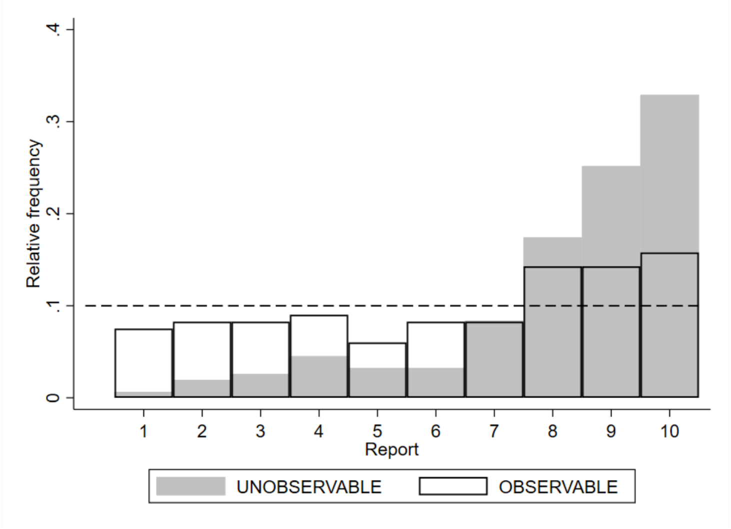
Figure 6 shows the distribution of reports in the UNOBSERVABLE and OBSERVABLE treatments. The dashed line in the figure indicates that in both treatments the truthful probability of drawing each state is 10 percent.

Finding 8 *Introducing observability has a strong and significant effect on the distribution of reports.*

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.

[30] 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.

Figure 6: Effect of changing the observability of states



Reports in the UNOBSERVABLE treatment are considerably higher than in the OBSERVABLE treatment (p < 0.001 OLS with robust SE; p < 0.001 Kolmogorov-Smirnov test; p < 0.001, Wilcoxon rank-sum test; see Kajackaite and Gneezy (2017) for a similar result).

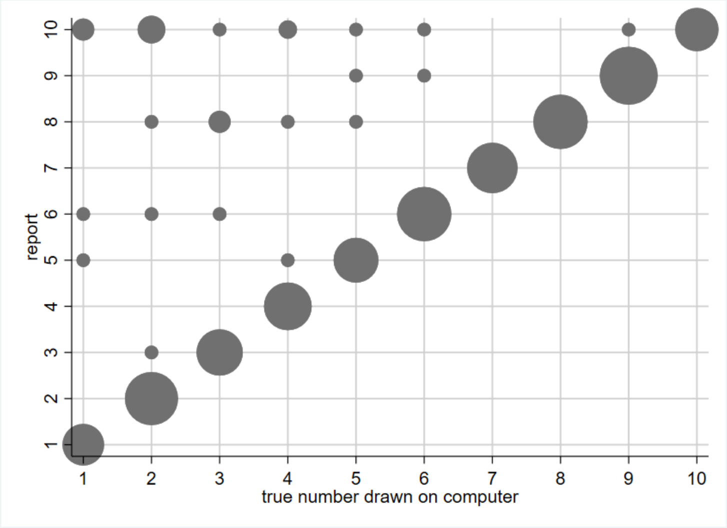
This result also demonstrates that it would be misleading to rely on evidence from settings in which the true state is observable by the researcher if one is actually interested in understanding a setting in which the true state is truly unobservable.

We can also use the OBSERVABLE treatment to examine our prediction about the existence of downward lying when the state is observable (Definition 4). Importantly, we may not have the same result in a setting where the true state is unobservable (see Table 2).

Finding 9 *There is no downward lying when the true state is observable.*

Figure 7 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 7: 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.

4.1 Overall Result of the Falsification Exercise

Recall that our four empirical tests, in addition to the meta study, 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 $\hat{g}$-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 do not). Taking all evidence together we find the following:

Finding 10 *Only the Reputation for Honesty + LC and the LC-Reputation models cannot be falsified by our data.*

Table 2 summarizes the predictions of all models. The two models that cannot be falsified by our data, Reputation for Honesty + LC and LC-Reputation, combine a preference for being honest with a preference for being seen as honest. In Reputation for Honesty + 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^{LC}.

All other models fail at least one of the four tests. Looking at Table 2, 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 Conformity in LC 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, as we discuss in Appendix C, several popular models, like the standard model and models that assume that subjects only care about their reputation for having been honest, cannot even explain the findings of the meta study (and also fail our new tests).

We find no significant effect of a change in beliefs, i.e., $\hat{g}$-invariance. As we discussed in Section 3.4, our study is sufficiently powered to detect treatment differences implied by reasonably parametrized versions of the social comparison models, e.g., Conformity in LC. We cannot, however, rule out (small) positive or negative effects of a change in beliefs. Regardless of whether our $\hat{G}$ treatments have enough power or not, even if we interpreted our data on this test as inconclusive and thus disregard the $\hat{g}$-invariance result, we can still reject all the social comparisons models because they fail at least one other experimental test.

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

4.2 A Calibrated Utility Function

In order to demonstrate how one of the non-falsified models, the Reputation for Honesty + LC model (Section 2.2.3), can quantitatively match both the data from the meta study and from our new experiments, we calibrate a simple, linear functional form. Our calibration is not intended to suggest that the functional form presented here, along with our choice of H, best matches the data. Instead, we view this as a demonstration that even quite simple and tractable assumptions generate equilibria that allow us to capture many of the important features of the data. Enriching the model further will only improve the fit. We suggest the following utility function which we call "Calibrated Reputation for Honesty + LC":

$$\phi(r, c(r, \omega), \Lambda(r); \theta^{RH}) = r - c\mathbb{I}_{\omega \neq r} - \theta^{RH}\Lambda(r)$$

As before, r is the report, ω the true state and $\Lambda(r)$ the fraction of liars at r. c is a fixed cost of lying and $\mathbb{I}_{\omega \neq r}$ is an indicator function of whether an individual lied. We suppose all individuals experience the same fixed cost of lying (this utility function is thus a limit case of the Reputation for Honesty + LC model). The individual-specific weight on reputation, θ^{RH}, is drawn from a uniform distribution on $[0, \kappa^{RH}]$. The average θ^{RH} is thus $\kappa^{RH}/2$. Additional details of the calibration are in Appendix H.2.[31]

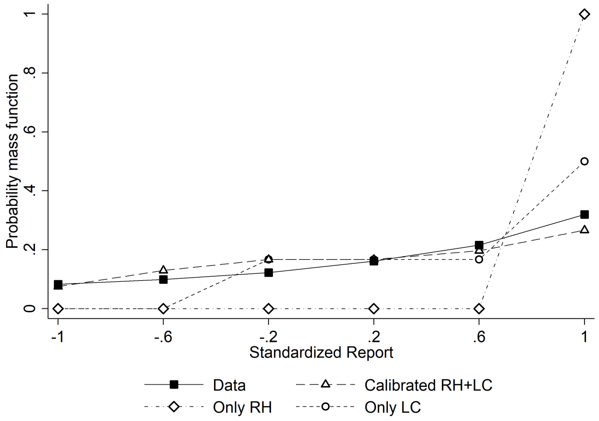
We calibrate the model to match the leading example in the literature, a simple die-roll setting, i.e., a uniform distribution F over six possible states with payoffs ranging from 1 to 6, where the audience cannot observe the state. We set $c = 3$ and $\kappa = 12$. We find that in the equilibrium, no individual lies down. Moreover, $\Lambda(r_i) = 0$ for $i \leq 4$, $\Lambda(r_5) \approx 0.15$ and $\Lambda(r_6) \approx 0.37$. We find a reporting distribution similar to that found in our meta study: $g(r_1) \approx 0.07$, $g(r_2) \approx 0.13$, $g(r_3) \approx 0.17$, $g(r_4) \approx 0.17$, $g(r_5) \approx 0.20$ and $g(r_6) \approx 0.27$. Figure 8 compares the predicted reporting distribution of this calibrated model to the data. The fit is quite good, in particular given the simple functional form, and the model matches all four findings of the meta study.

It also matches up with our experimental findings. In a setting where the state is observ-

[31] In concurrent work, Khalmetski and Sliwka (forthcoming) and Gneezy et al. (2018) discuss another limit case of the Reputation for Honesty + LC model, where all individuals face the same reputational cost, but vary in the LC component of utility. Such utility functions can also be calibrated to match both the meta study data and our new experiments.

able, the model predicts no downward lying, as in our data (this is true for all Reputation for Honesty + LC utility functions), and much more truth-telling. Under observability, all liars report the maximal report, similar to our data.

Figure 8: Calibrated Reputation for Honesty + LC



The model also generates the large amount of drawing in we observe. We consider two states like in our F treatments, and in order to keep the payoff scale the same as the previous calibration, we suppose they pay 1 and 6. When $f(\omega_1) = 0.4$, the equilibrium features no lying down and so $\Lambda(r_1) = 0$. Moreover, $\Lambda(r_6) \approx 0.28$ and the share of low reports is $g(r_1) \approx 0.16$. When $f(\omega_1) = 0.9$, we find two equilibria. One of the equilibria features no lying down, and in this case $\Lambda(r_6) \approx 0.69$ and $g(r_1) \approx 0.68$. The other equilibrium features lying down; here $\Lambda(r_1) \approx 0.10$, $\Lambda(r_6) \approx 0.91$ and $g(r_1) \approx 0.80$. Thus, in the last equilibrium, approximately 8 out of every 10 individuals who draw the high state give the low report. For comparison, our experiments yield $g(r_1) = 0.13$ and $g(r_1) = 0.55$, respectively. Regardless of which of these two equilibria is selected, we observe significant amounts of drawing in. Moreover, the model can generate almost any behavior in our $\hat{G}$ treatments, because those treatments do not pin down the belief about H (and thus $\Lambda(r)$, on which utility in the model depends). Depending on the new beliefs, aversion, $\hat{g}$-invariance or affinity could result; as the new belief could either imply a positive, no, or negative change in the gap between $\Lambda(r_6)$ and $\Lambda(r_1)$ (see the Reputation for Honesty + LC part of the proof of Proposition 2 for details).

Both components of the utility function are important. In Figure 8, we also plot the predicted reporting distributions for the utility function when we shut down the LC or the RH part. The Only-RH model is far away from the data. The Only-LC model is closer, but this model does not generate drawing in or o-shift.[32]

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" findings 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 a preference for being seen as honest and a preference for being honest are the main motivations for truth-telling. 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.[33]

Three concurrent papers also present models that incorporate a desire to appear honest in the utility function. The utility functions proposed by Khalmetski and Sliwka (forthcoming) and Gneezy et al. (2018) are similar in spirit to our Reputation for Honesty + LC model. Both papers combine a desire to appear honest with a desire to be honest. Khalmetski and Sliwka (forthcoming) show that a calibrated version of their model reproduces the data patterns observed in the FFH paradigm. Similar to two of our new tests, Gneezy et al. (2018) present experiments that manipulate the true distribution of the states as well as the observability of the state, with similar results to our tests. Taken together, the results of these two studies are in line with the two non-falsified models we propose that also combine lying costs and reputational costs. In another concurrent paper, M. Dufwenberg and M. A. Dufwenberg (2018) present a different, more nuanced formalization of the desire to appear honest, in particular they assume that individuals care about the beliefs that an audience has about the degree

[32] In the Only-LC model, individuals who draw ω_3 are indifferent between reporting r_3 and r_6. We suppose for the figure that they say r_3. Shifting these to r_6 only worsens the fit.

[33] 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.

of over-reporting (rather than the simple chance of being a liar). M. Dufwenberg and M. A. Dufwenberg (2018) show that this model can explain the results of the original Fischbacher and Föllmi-Heusi (2013) setup (six-sided die roll). Future research could investigate whether reputational concerns regarding honesty are more often captured by the assumptions in the models of Khalmetski and Sliwka (forthcoming), Gneezy et al. (2018) and our paper or by the M. Dufwenberg and M. A. Dufwenberg (2018) assumption of perceived cheating aversion.

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 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 (e.g., Bø et al. 2015).

There are at least four potential caveats for these policy implications. First, we would not 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 0.234. 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 they 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 (Bénabou et al. 2018, Mazar et al. 2008).

References

Abeler, Johannes (2015). “A reporting experiment with Chinese professionals”. *mimeo.*

Abeler, Johannes, Anke Becker, and Armin Falk (2014). “Representative evidence on lying costs”. *Journal of Public Economics* 113, pp. 96–104.

Abeler, Johannes and Daniele Nosenzo (2015). “Lying and other preferences”. *mimeo.*

Akerlof, George (1983). “Loyalty filters”. *American Economic Review* 73.1, pp. 54–63.

Allingham, Michael and Agnar Sandmo (1972). “Income tax evasion: A theoretical analysis”. *Journal of Public Economic* 1, pp. 323–338.

Amir, Amitai, Tehila Kogut, and Yoella Bereby-Meyer (2016). “Careful cheating: people cheat groups rather than individuals”. *Frontiers in Psychology* 7.

Antony, Markus, Holger Gerhardt, and Armin Falk (2016). “The Impact of Food and Water Deprivation on Economic Decision Making”. *mimeo.*

Arbel, Yuval, Ronen Bar-El, Erez Siniver, and Yossef Tobol (2014). “Roll a die and tell a lie–What affects honesty?” *Journal of Economic Behavior & Organization* 107, pp. 153–172.

Ariely, Dan, Ximena Garcia-Rada, Lars Hornuf, and Heather Mann (2014). “The (true) legacy of two really existing economic systems”. *University of Munich Discussion Paper.*

Aydogan, Gökhan, Andrea Jobst, Kimberlee D’Ardenne, Norbert Müller, and Martin Kocher (2017). “The detrimental effects of oxytocin-induced conformity on dishonesty in competition”. *Psychological Science* 28.6, pp. 751–759.

Banerjee, Ritwik, Nabanita Datta Gupta, and Marie Claire Villeval (2018). “The spillover effects of affirmative action on competitiveness and unethical behavior”. *European Economic Review* 101, pp. 567–604.

Barfort, Sebastian, Nikolaj Harmon, Frederik Hjorth, and Asmus Leth Olsen (2015). “Dishonesty and selection into public service in Denmark: Who runs the world’s least corrupt public service?” *University of Copenhagen Discussion Paper.*

Basic, Zvonimir, Armin Falk, and Simone Quercia (2016). “The influence of self and social image concerns on lying”. *mimeo.*

Batson, Daniel, Diane Kobrynowicz, Jessica Dinnerstein, Hannah Kampf, and Angela Wilson (1997). “In a very different voice: Unmasking moral hypocrisy.” *Journal of Personality and Social Psychology* 72.6, p. 1335.

Battigalli, Pierpaolo and Martin Dufwenberg (2009). “Dynamic psychological games”. *Journal of Economic Theory* 144.1, pp. 1–35.

Beck, Tobias, Christoph Bühren, Björn Frank, and Elina Khachatryan (2018). “Can Honesty Oaths, Peer Interaction, or Monitoring Mitigate Lying?” *Journal of Business Ethics.*

Bénabou, Roland, Armin Falk, and Jean Tirole (2018). “Narratives, Imperatives, and Moral Reasoning”. *CEPR DP 13056.*

Bénabou, Roland and Jean Tirole (2006). “Incentives and Prosocial Behavior”. *American Economic Review* 96.5, pp. 1652–1678.

Blanco, Mariana and Juan-Camilo Cárdenas (2015). “Honesty after a labor relationship”. *Universidad del Rosario Discussion Paper* 2015-37.

Bø, Erlend, Joel Slemrod, and Thor Thoresen (2015). “Taxes on the internet: Deterrence effects of public disclosure”. *American Economic Journal: Economic Policy* 7.1, pp. 36–62.

Braun, Susanne and Lars Hornuf (2015). “Leadership and persistency in spontaneous dishonesty”. *IAAEU Discussion Paper.*

Bryan, Christopher, Gabrielle Adams, and Benoit Monin (2013). “When cheating would make you a cheater: Implicating the self prevents unethical behavior.” *Journal of Experimental Psychology: General* 142.4, p. 1001.

Bucciol, Alessandro and Marco Piovesan (2011). “Luck or cheating? A field experiment on honesty with children”. *Journal of Economic Psychology* 32.1, pp. 73–78.

Cadsby, Bram, Ninghua Du, and Fei Song (2016). “In-group favoritism and moral decision-making”. *Journal of Economic Behavior and Organization* 128, pp. 59–71.

Cappelen, Alexander W., Odd-Helge Fjeldstad, Donald Mmari, Ingrid H. Sjursen, and Bertil Tungodden (2016). “Managing the resource curse: A survey experiment on expectations about gas revenues in Tanzania”. *mimeo.*

Cardenas, Juan Camilo and Jeffrey Carpenter (2008). “Behavioural development economics: Lessons from field labs in the developing world”. *Journal of Development Studies* 44.3, pp. 311–338.

Charness, Gary, Celia Blanco-Jimenez, Lara Ezquerra, and Ismael Rodriguez-Lara (forthcoming). "Cheating, incentives, and money manipulation". *Experimental Economics*.

Charness, Gary and Martin Dufwenberg (2006). "Promises and partnership". *Econometrica* 74.6, pp. 1579–1601.

Chytilova, Julie and Vaclav Korbel (2014). "Individual and group cheating behavior: a field experiment with adolescents". *IES Working Paper*.

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez (2014). "Smug Alert! Exploring self-licensing behavior in a cheating game". *Economics Letters* 123.2, pp. 191–194.

Cohn, Alain, Ernst Fehr, and Michel André Maréchal (2014). "Business culture and dishonesty in the banking industry". *Nature* 516.7529, pp. 86–89.

Cohn, Alain, Tobias Gesche, and Michel Maréchal (2018). "Honesty in the Digital Age". *University of Zurich Working Paper*.

Cohn, Alain and Michel Maréchal (forthcoming). "Laboratory Measure of Cheating Predicts School Misconduct". *Economic Journal*.

Cohn, Alain, Michel André Maréchal, and Thomas Noll (2015). "Bad boys: How criminal identity salience affects rule violation". *Review of Economic Studies* 82.4, pp. 1289–1308.

Conrads, Julian, Mischa Ellenberger, Bernd Irlenbusch, Elia Ohms, Rainer Rilke, and Gari Walkowitz (2017). "Team Goal Incentives and Individual Lying Behavior". *WHU Discussion Paper*.

Conrads, Julian, Bernd Irlenbusch, Rainer Michael Rilke, Anne Schielke, and Gari Walkowitz (2014). "Honesty in tournaments". *Economics Letters* 123.1, pp. 90–93.

Conrads, Julian, Bernd Irlenbusch, Rainer Michael Rilke, and Gari Walkowitz (2013). "Lying and team incentives". *Journal of Economic Psychology* 34, pp. 1–7.

Conrads, Julian and Sebastian Lotz (2015). "The effect of communication channels on dishonest behavior". *Journal of Behavioral and Experimental Economics* 58, pp. 88–93.

Crawford, Vincent and Joel Sobel (1982). "Strategic information transmission". *Econometrica* 50.6, pp. 1431–1451.

Dai, Zhixin, Fabio Galeotti, and Marie Claire Villeval (2018). "Cheating in the Lab Predicts Fraud in the Field: An Experiment in Public Transportation". *Management Science* 64.3, pp. 1081–1100.

Dato, Simon and Petra Nieken (2016). “Compensation and Honesty: Gender Differences in Lying”. *Beiträge zur Jahrestagung des Vereins für Socialpolitik.*

DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao (2016). “Voting to tell others”. *The Review of Economic Studies* 84.1, pp. 143–181.

Di Falco, Salvatore, Brice Magdalou, David Masclet, Marie Claire Villeval, and Marc Willinger (2016). “Can transparency of information reduce embezzlement? Experimental Evidence from Tanzania”. *IZA Working Paper.*

Dieckmann, Anja, Veronika Grimm, Matthias Unfried, Verena Utikal, and Lorenzo Valmasoni (2016). “On trust in honesty and volunteering among Europeans: Cross-country evidence on perceptions and behavior”. *European Economic Review* 90, pp. 225–253.

Diekmann, Andreas, Wojtek Przepiorka, and Heiko Rauhut (2015). “Lifting the veil of ignorance: An experiment on the contagiousness of norm violations”. *Rationality and Society* 27.3, pp. 309–333.

Djawadi, Behnud Mir and René Fahr (2015). ““. . . and they are really lying”: Clean evidence on the pervasiveness of cheating in professional contexts from a field experiment”. *Journal of Economic Psychology* 48, pp. 48–59.

Drupp, Moritz, Menusch Khadjavi, and Martin Quaas (2016). “Truth-Telling and the Regulator: Evidence from a Field Experiment with Commercial Fishermen”. *Kiel Working Paper.*

Duch, Raymond and Hector Solaz (2016). “Who Cheats: Experimental Evidence from the Lab”. *Discussion Paper, CESS, Nuffield College, University of Oxford.*

Dufwenberg, Martin and Martin A. Dufwenberg (2018). “Lies in Disguise – A Theoretical Analysis of Cheating”. *Journal of Economic Theory* 175, pp. 248–264.

Effron, Daniel, Christopher Bryan, and Keith Murnighan (2015). “Cheating at the end to avoid regret.” *Journal of personality and social psychology* 109.3, p. 395.

Ellingsen, Tore and Magnus Johannesson (2004). “Promises, threats and fairness”. *The Economic Journal* 114.495, pp. 397–420.

Ellingsen, Tore and Robert Östling (2010). “When does communication improve coordination?” *American Economic Review* 100.4, pp. 1695–1724.

Engel, Christoph (2011). “Dictator games: A meta study”. *Experimental Economics* 14.4, pp. 583–610.

Fischbacher, Urs (2007). “z-Tree: Zurich toolbox for ready-made economic experiments”. *Experimental Economics* 10.2, pp. 171–178. (Visited on 03/25/2009).

Fischbacher, Urs and Franziska Föllmi-Heusi (2013). “Lies in disguise—an experimental study on cheating”. *Journal of the European Economic Association* 11.3, pp. 525–547.

Foerster, Anna, Roland Pfister, Constantin Schmidts, David Dignath, and Wilfried Kunde (2013). “Honesty saves time (and justifications)”. *Frontiers in Psychology* 4.

Fosgaard, Toke Reinholt (2013). “Asymmetric default bias in dishonesty–how defaults work but only when in one’s favor”. *University of Copenhagen Discussion Paper.*

Fosgaard, Toke Reinholt, Lars Gaarn Hansen, and Marco Piovesan (2013). “Separating Will from Grace: An experiment on conformity and awareness in cheating”. *Journal of Economic Behavior & Organization* 93, pp. 279–284.

Gächter, Simon and Jonathan Schulz (2016). “Lying and beliefs”. *mimeo.*

Gächter, Simon and Jonathan F Schulz (2016a). *Data from: Intrinsic honesty and the prevalence of rule violations across societies.* `http://dx.doi.org/10.5061/dryad.9k358`. Dryad Digital Repository.

— (2016b). “Intrinsic honesty and the prevalence of rule violations across societies”. *Nature* 531, pp. 496–499.

Garbarino, Ellen, Robert Slonim, and Marie Claire Villeval (forthcoming). “Loss Aversion and lying behavior: Theory, estimation and empirical evidence”. *Journal of Economic Behavior and Organization.*

Geanakoplos, John, David Pearce, and Ennio Stacchetti (1989). “Psychological Games and Sequential Rationality”. *Games and Economic Behavior* 1, pp. 60–79.

Gibson, Rajna, Carmen Tanner, and Alexander Wagner (2013). “Preferences for truthfulness: Heterogeneity among and within individuals”. *American Economic Review* 103, pp. 532–548.

Gill, David, Victoria Prowse, and Michael Vlassopoulos (2013). “Cheating in the workplace: An experimental study of the impact of bonuses and productivity”. *Journal of Economic Behavior & Organization* 96, pp. 120–134.

Gino, Francesca and Dan Ariely (2012). “The dark side of creativity: original thinkers can be more dishonest.” *Journal of Personality and Social Psychology* 102.3, p. 445.

Gneezy, Uri (2005). “Deception: The role of consequences”. *American Economic Review* 95.1, pp. 384–394.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel (2018). “Lying Aversion and the Size of the Lie”. *American Economic Review* 108.2, pp. 419–453.

Gneezy, Uri, Bettina Rockenbach, and Marta Serra-Garcia (2013). “Measuring lying aversion”. *Journal of Economic Behavior & Organization* 93, pp. 293–300.

Greiner, Ben (2015). “Subject pool recruitment procedures: organizing experiments with ORSEE”. *Journal of the Economic Science Association* 1.1, pp. 114–125.

Grigorieff, Alexis and Christopher Roth (2016). “How Does Economic Status Affect Social Preferences? Representative Evidence From a Survey Experiment”. *mimeo.*

Halevy, Rony, Shaul Shalvi, and Bruno Verschuere (2014). “Being Honest About Dishonesty: Correlating Self-Reports and Actual Lying”. *Human Communication Research* 40.1, pp. 54–72.

Hanna, Rema and Shing-Yi Wang (2017). “Dishonesty and selection into public service: Evidence from India”. *American Economic Journal: Economic Policy* 9.3, pp. 262–90.

Hao, Li and Daniel Houser (2017). “Perceptions, intentions, and cheating”. *Journal of Economic Behavior & Organization* 133, pp. 52–73.

Harless, David and Colin Camerer (1994). “The predictive utility of generalized expected utility theories”. *Econometrica* 62.6, pp. 1251–1289.

Heldring, Leander (2016). “Violence and the State: Evidence from Rwanda’s ‘Decade of Atrocities’”. *mimeo.*

Hilbig, Benjamin and Corinna Hessler (2013). “What lies beneath: How the distance between truth and lie drives dishonesty”. *Journal of Experimental Social Psychology* 49.2, pp. 263–266.

Hilbig, Benjamin and Ingo Zettler (2015). “When the cat’s away, some mice will play: A basic trait account of dishonest behavior”. *Journal of Research in Personality* 57, pp. 72–88.

Houser, Daniel, John A List, Marco Piovesan, Anya Samek, and Joachim Winter (2016). “Dishonesty: From parents to children”. *European Economic Review* 82, pp. 242–254.

Houser, Daniel, Stefan Vetter, and Joachim Winter (2012). “Fairness and cheating”. *European Economic Review* 56.8, pp. 1645–1655.

Hruschka, Daniel, Charles Efferson, Ting Jiang, Ashlan Falletta-Cowden, Sveinn Sigurdsson, Rita McNamara, Madeline Sands, Shirajum Munira, Edward Slingerland, and Joseph Henrich (2014). “Impartial institutions, pathogen stress and the expanding social network”. *Human Nature* 25.4, pp. 567–579.

Hugh-Jones, David (2016). “Honesty, beliefs about honesty, and economic growth in 15 countries”. *Journal of Economic Behavior & Organization* 127, pp. 99–114.

Jacobsen, Catrine and Marco Piovesan (2016). “Tax me if you can: An artefactual field experiment on dishonesty”. *Journal of Economic Behavior and Organization* 124, pp. 7–14.

Jiang, Ting (2013). “Cheating in mind games: The subtlety of rules matters”. *Journal of Economic Behavior & Organization* 93, pp. 328–336.

— (2015). “Other-Regarding Preferences and Other-Regarding Cheating – Experimental Evidence from China, Italy, Japan and the Netherlands”. *SSRN Working Paper*.

Johnson, Noel and Alexandra Mislin (2011). “Trust games: A meta-analysis”. *Journal of Economic Psychology* 32.5, pp. 865–889.

Kajackaite, Agne and Uri Gneezy (2017). “Incentives and cheating”. *Games and Economic Behavior* 102, pp. 433–444.

Kartik, Navin (2009). “Strategic communication with lying costs”. *Review of Economic Studies* 76.4, pp. 1359–1395.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani (2007). “Credulity, lies, and costly talk”. *Journal of Economic Theory* 134.1, pp. 93–116.

Kartik, Navin, Olivier Tercieux, and Richard Holden (2014). “Simple mechanisms and preferences for honesty”. *Games and Economic Behavior* 83, pp. 284–290.

Khalmetski, Kiryl and Dirk Sliwka (forthcoming). “Disguising Lies – Image Concerns and Partial Lying in Cheating Games”. *American Economic Journal: Microeconomics*.

Kroher, Martina and Tobias Wolbring (2015). “Social control, social learning, and cheating: Evidence from lab and online experiments on dishonesty”. *Social Science Research* 53, pp. 311–324.

Lowes, Sara, Nathan Nunn, James A Robinson, and Jonathan L Weigel (2017). “The evolution of culture and institutions: Evidence from the Kuba Kingdom”. *Econometrica* 85.4, pp. 1065–1091.

Ma, Ching-To Albert and Thomas McGuire (1997). “Optimal health insurance and provider payment”. *American Economic Review* 87.4, pp. 685–704.

Maggian, Valeria and Natalia Montinari (2017). “The spillover effects of gender quotas on dishonesty”. *Economics Letters* 159, pp. 33–36.

Mann, Heather, Ximena Garcia-Rada, Lars Hornuf, Juan Tafurt, and Dan Ariely (2016). “Cut from the Same Cloth: Surprisingly Honest Individuals Across Cultures”. *Journal of Cross-Cultural Psychology* 47.6, pp. 858–874.

Matsushima, Hitoshi (2008). “Role of honesty in full implementation”. *Journal of Economic Theory* 139.1, pp. 353–359.

Mazar, Nina, On Amir, and Dan Ariely (2008). “The dishonesty of honest people: A theory of self-concept maintenance”. *Journal of Marketing Research* 45.6, pp. 633–644.

Meub, Lukas, Till Proeger, Tim Schneider, and Kilian Bizer (2016). “The victim matters–Experimental evidence on lying, moral costs and moral cleansing”. *Applied Economics Letters* 23.16, pp. 1162–1167.

Muehlheusser, Gerd, Andreas Roider, and Niklas Wallmeier (2015). “Gender differences in honesty: Groups versus individuals”. *Economics Letters* 128, pp. 25–29.

Muñoz-Izquierdo, Nora, Beatriz Gil-Gómez de Liaño, Francisco Daniel Rin-Sánchez, and David Pascual-Ezama (2014). “Economists: cheaters with altruistic instincts”. *MPRA Discussion Paper.*

Oosterbeek, Hessel, Randolph Sloof, and Gijs Van De Kuilen (2004). “Cultural differences in ultimatum game experiments: Evidence from a meta-analysis”. *Experimental Economics* 7.2, pp. 171–188.

Pascual-Ezama, David et al. (2015). “Context-dependent cheating: Experimental evidence from 16 countries”. *Journal of Economic Behavior & Organization* 116, pp. 379–386.

Peer, Eyal, Alessandro Acquisti, and Shaul Shalvi (2014). “’I cheated, but only a little’: Partial confessions to unethical behavior.” *Journal of Personality and Social Psychology* 106.2, pp. 202–217.

Ploner, Matteo and Tobias Regner (2013). “Self-image and moral balancing: An experimental analysis”. *Journal of Economic Behavior & Organization* 93, pp. 374–383.

Popper, Karl (1934). *Logik der Forschung.* Vienna: Julius Springer.

Potters, Jan and Jan Stoop (2016). “Do cheaters in the lab also cheat in the field?” *European Economic Review* 87, pp. 26–33.

Rauhut, Heiko (2013). “Beliefs about Lying and Spreading of Dishonesty: Undetected Lies and Their Constructive and Destructive Social Dynamics in Dice Experiments”. *PLoS One* 8.11.

Ruedy, Nicole and Maurice Schweitzer (2010). “In the moment: The effect of mindfulness on ethical decision making”. *Journal of Business Ethics* 95.1, pp. 73–87.

Ruffle, Bradley and Yossef Tobol (2014). “Honest on Mondays: Honesty and the temporal separation between decisions and payoffs”. *European Economic Review* 65, pp. 126–135.

Sánchez-Pagés, Santiago and Marc Vorsatz (2009). “Enjoy the silence: an experiment on truth-telling”. *Experimental Economics* 12.2, pp. 220–241.

Sandholm, William (2015). “Population Games and Deterministic Evolutionary Dynamics”. *Handbook of Game Theory.* Ed. by Peyton Young and Shmuel Zamir. Vol. 4. Elsevier, pp. 703–775.

Schindler, Simon and Stefan Pfattheicher (2017). “The frame of the game: Loss-framing increases dishonest behavior”. *Journal of Experimental Social Psychology* 69, pp. 172–177.

Serra-Garcia, Marta, Eric Van Damme, and Jan Potters (2011). “Hiding an inconvenient truth: Lies and vagueness”. *Games and Economic Behavior* 73.1, pp. 244–261.

Shalvi, Shaul (2012). “Dishonestly increasing the likelihood of winning”. *Judgment and Decision Making* 7.3, p. 292.

Shalvi, Shaul, Jason Dana, Michel Handgraaf, and Carsten De Dreu (2011). “Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior”. *Organizational Behavior and Human Decision Processes* 115.2, pp. 181–190.

Shalvi, Shaul and Carsten De Dreu (2014). “Oxytocin promotes group-serving dishonesty”. *Proceedings of the National Academy of Sciences* 111.15, pp. 5503–5507.

Shalvi, Shaul, Ori Eldar, and Yoella Bereby-Meyer (2012). “Honesty requires time (and lack of justifications)”. *Psychological Science* 23.10, pp. 1264–1270.

Shalvi, Shaul and David Leiser (2013). “Moral firmness”. *Journal of Economic Behavior & Organization* 93, pp. 400–407.

Shen, Qiang, Meijun Teo, Eyal Winter, Einav Hart, Soo H Chew, and Richard P Ebstein (2016). “To cheat or not to cheat: Tryptophan hydroxylase 2 SNP variants contribute to dishonest behavior”. *Frontiers in Behavioral Neuroscience* 10.

Škoda, Samuel (2013). “Effort and cheating behavior: An experiment”. *mimeo*.

Suri, Siddharth, Daniel Goldstein, and Winter Mason (2011). “Honesty in an Online Labor Market”. *Human Computation*.

Thielmann, Isabel, Benjamin E Hilbig, Ingo Zettler, and Morten Moshagen (2017). “On measuring the sixth basic personality dimension: A comparison between HEXACO Honesty-Humility and Big Six Honesty-Propriety”. *Assessment* 24.8, pp. 1024–1036.

Tversky, Amos and Daniel Kahneman (1974). “Judgment under uncertainty: Heuristics and biases”. *Science* 185.4157, pp. 1124–1131.

Utikal, Verena and Urs Fischbacher (2013). “Disadvantageous lies in individual decisions”. *Journal of Economic Behavior & Organization* 85, pp. 108–111.

Vanberg, Christoph (2008). “Why do people keep their promises? An experimental test of two explanations”. *Econometrica* 76.6, pp. 1467–1480.

Warner, Stanley (1965). “Randomized response: A survey technique for eliminating evasive answer bias”. *Journal of the American Statistical Association* 60.309, pp. 63–69.

Waubert De Puiseau, Berenike and Andreas Glöckner (2012). “Investigating cheating behavior in students compared to the general public”. *mimeo*.

Weibull, Jörgen and Edgar Villa (2005). “Crime, punishment and social norms”. *SSE/EFI Discussion Paper*.

Weisel, Ori and Shaul Shalvi (2015). “The collaborative roots of corruption”. *Proceedings of the National Academy of Sciences* 112.34, pp. 10651–10656.

Weizsäcker, Georg (2010). “Do We Follow Others When We Should? A Simple Test of Rational Expectations”. *American Economic Review* 100.5, pp. 2340–60.

Wibral, Matthias, Thomas Dohmen, Dietrich Klingmüller, Bernd Weber, and Armin Falk (2012). “Testosterone Administration Reduces Lying in Men”. *PLoS One* 7.10.

Zettler, Ingo, Benjamin Hilbig, Morten Moshagen, and Reinout de Vries (2015). “Dishonest responding or true virtue? A behavioral test of impression management”. *Personality and Individual Differences* 81, pp. 107–111.

Zimerman, Liora, Shaul Shalvi, Yoella Bereby-Meyer, et al. (2014). “Self-reported ethical risk taking tendencies predict actual dishonesty”. *Judgment and Decision Making* 9.1, pp. 58–64.

Online Appendices

This document contains the online appendices for the paper "Preferences for truth-telling" by Johannes Abeler, Daniele Nosenzo and Collin Raymond.

- Appendix A contains further results of the meta study.
- Appendix B presents and derives predictions for those models listed in Table 2 of the main body of the paper that were not discussed in the body of the paper.
- Appendix C discusses some prominent models that are discussed in the literature but that cannot explain the findings of the meta study and are thus not discussed in the main body of the paper.
- Appendix D contains the proofs for the predictions of the models presented in Section 2 in the main body of the paper.
- Appendix E explores how predictions would change if we altered the assumptions regarding the distribution H of individual-level parameters $\vec{\theta}$.
- Appendix F presents two additional sets of experiments that we conducted to test specific predictions of some of the models considered in the paper.
- Appendix G contains the instructions for the lab experiments.
- Appendix H explains the details of the calibrations in Section 4 in the body of the paper.

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 M. 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. (2014) or d'Adda et al. (2017). 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), a treatment of Gneezy et al. (2018) and two of our treatments reported in this paper.[34]

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 assumption 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.[35]

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 specifications with study fixed effects are in Table A.2, columns 3 to 8 (in column 9, we also report the within-study difference for students vs. non-students even though being a

[34] We adjust the distribution of standardized reports of experiments with asymmetric distributions and two states such that the average standardized report is comparable to the one of symmetric distributions.

[35] We restrict explanatory variables in this way since otherwise any treatment fixed effect could be an explanatory variable. Given that we include 429 treatments this would become unwieldy.

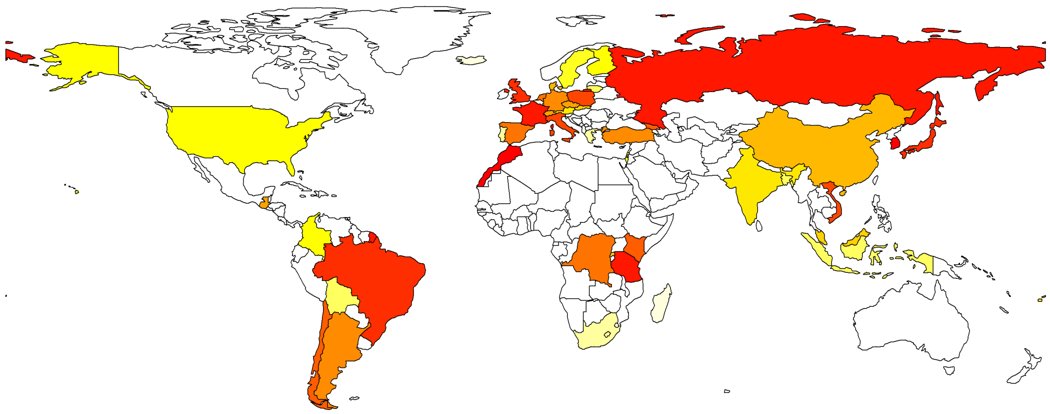
student is not randomly allocated).

Table A.1: Meta study: descriptive statistics

	Mean	# Subjects
Treatment-level variables		
Maximal payoff from misreporting (in 2015 USD)	4.480	44390
1 if repeated	0.244	44390
1 if online/telephone	0.273	44390
1 if control rolls suggested	0.283	44390
1 if reporting about state of mind	0.167	44390
1 if info about behavior of other subjects available	0.011	44390
1 if report reduces payoff of another subject	0.032	44390
1 if student subjects	0.577	44390
Year experiment conducted	2013.460	44390
Author affiliation		
1 if economics	0.758	44390
1 if psychology	0.212	44390
1 if sociology/anthropology	0.030	44390
Method of randomization		
1 if coin toss	0.421	44390
1 if die roll	0.529	44390
1 if draw from urn	0.050	44390
True distribution		
1 if two outcomes non-uniform	0.122	44390
1 if two outcomes uniform	0.358	44390
1 if other uniform	0.370	44390
1 if bell shaped	0.150	44390
Individual-level variables		
1 if female	0.478	22944
Age	29.652	16205
Field of study		
1 if economics/management student	0.242	5284
1 if psychology student	0.027	5284
1 if other student	0.731	5284
# Decisions	270616	
# Subjects	44390	
# Treatments	429	
# Studies	90	

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 two specifications could yield different estimates for three reasons: (i) cleaner identification in the within-study specification, (ii) publication bias, and (iii) treatment effect heterogeneity. First, if there are important omitted variables in the between-study specification, the estimated coefficients will be biased. Omitted variables are not an issue for the within-study specification. Second, we would expect that studies that do not find a significant treatment effect are less likely to get published and are thus less likely to be included in our meta study. This will bias upwards the coefficient in the within-study specification. The between-studies specification suffers much less from this publication bias as we collect information about variables which the original authors did not use for their publication decision. If publication bias is important, then our between-study specification should give a better estimate of the true coefficient than the within-study specification. Third, the within-study estimates only use data from studies that vary the parameter of interest directly, thus restricting the sample considerably. If there is treatment effect heterogeneity, we would expect the within-study estimate to differ from the between-study estimate. For example, the incentive level could have a stronger effect for student samples than for non-student samples. We find that treatment effect heterogeneity could indeed explain the difference between within-study and between-study coefficients.[36] If one is only interested in the average treatment effect,

[36] Take the incentive level coefficient as example. The between-study coefficient is -0.005 (see below for details, based on 429 treatments) and the within-study coefficient is 0.003 (based on 94 treatments). To test whether treatment effect heterogeneity could explain this difference, we take the entire sample, draw 94 treatments at

the between-study specification is thus preferable as it reports the average effect of a larger sample. Taken together, since we do not know with certainty how important the three reasons are, we can only say that both estimates are informative. We thus report results of regressions using both identification strategies. It turns out that in Table A.2, only one coefficient out of six is different from zero and has an opposite sign in the between- and within-study regressions.

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.[37] In the regressions relying on conditional independence, we also report a specification in column 2 which clusters on study to allow for dependencies within study. Independent of clustering, we weight one decision as one observation in all regressions.[38]

random and run the between-study specification on this subsample. We repeat this process 10000 times. We find that 28 percent of the between-study coefficients are larger than 0.003.

[37] 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.

[38] If we weight by subject, results are very similar. Only the overall average standardized report is then 0.321 instead of 0.234.

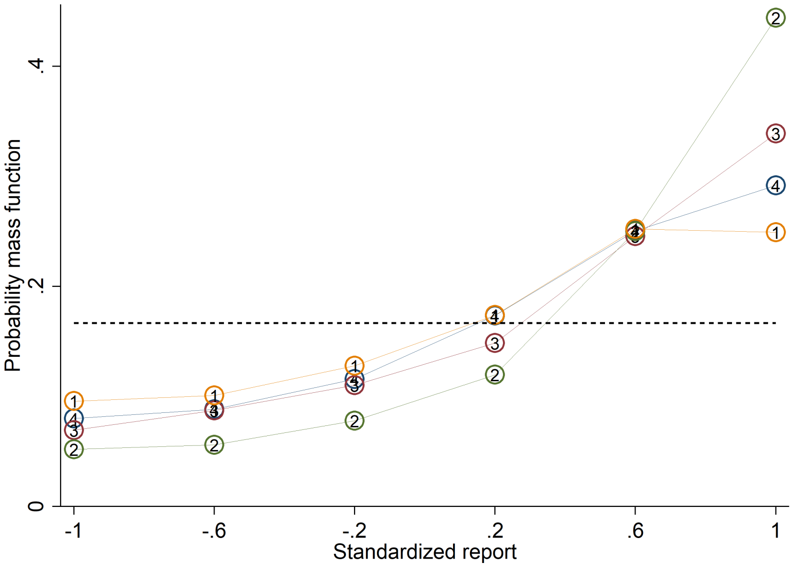
Table A.2: Regressions of treatment-level variables

Dependent variable: Standardized report	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Maximal payoff from misreporting	-0.005	-0.005	0.003						
	(0.001)	(0.003)	(0.001)						
1 if repeated	-0.118	-0.118							
	(0.012)	(0.051)							
1 if online/telephone	-0.006	-0.006		0.027					
	(0.012)	(0.051)		(0.032)					
1 if control rolls suggested	0.037	0.037			0.168				
	(0.013)	(0.049)			(0.052)				
1 if reporting about state of mind	0.051	0.051				0.173			
	(0.013)	(0.072)				(0.034)			
1 if info about behavior of other subjects available	0.043	0.043					0.046		
	(0.042)	(0.012)					(0.048)		
1 if report reduces payoff of another subject	-0.045	-0.045						-0.095	
	(0.014)	(0.050)						(0.017)	
1 if student subjects	0.095	0.095							0.100
	(0.009)	(0.042)							(0.026)
Year experiment conducted	-0.002	-0.002							
	(0.002)	(0.005)							
Additional controls	No	No	Yes	No	Yes	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	No	No	No	No	No	Yes
Study FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# Decisions	270616	270616	51335	1214	577	1200	24349	5510	91593
# Subjects	44390	44390	7896	906	577	1200	1642	3275	4792
# Treatments	429	429	94	12	14	16	22	42	40
# Studies	90	90	11	3	3	1	5	6	8
# Clusters	44213	90	7896	906	577	1200	1465	3275	4792

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.

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 1 and 2. An increase of the potential payoff by 1 USD changes the standardized report by -0.005. 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 (this represents about a third of the data set). 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. In almost all cases, higher states are reported more often than lower states, 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

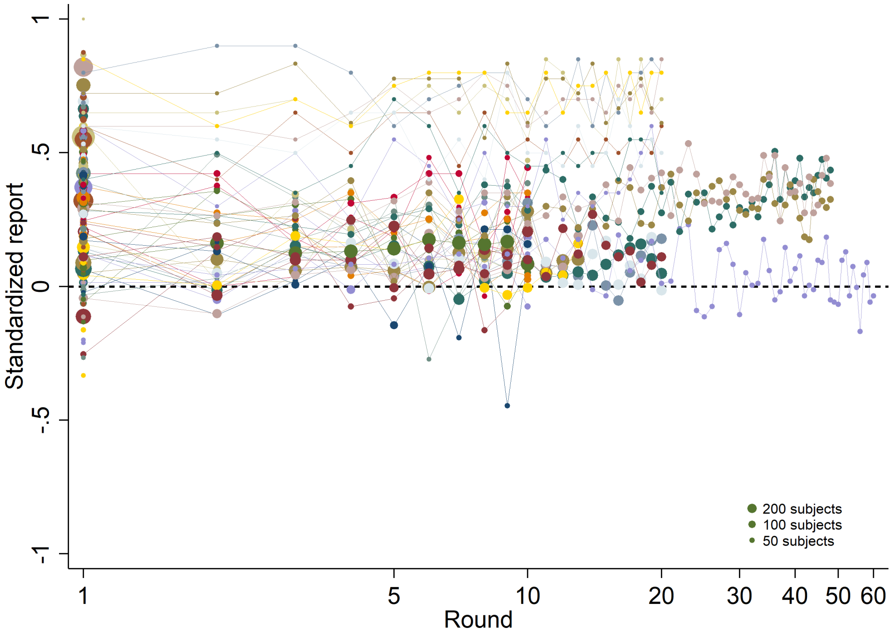


Notes: The figure depicts the distribution of reports for treatments that use a uniform distribution with six states and linear payoff increases. Treatments are collapsed into quartiles by the level of the maximal payoff from misreporting. The line marked by “1” is the distribution of the treatments 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: The regressions in Table A.2, columns 1 and 2, show that experiments with repeated reports induce on average markedly 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. Visually, there is no strong 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. Subjects in repeated experiments thus start lower than subjects in one-shot experiments and then slowly gravitate towards the level of one-shot behavior. This pattern 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.

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 that suggesting control rolls increases reports significantly (columns 1 and 5).[39]

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 between-study results in column 1 show that reporting about a state of mind leads to significantly higher reports. The one study which tested this difference directly (Kajackaite and Gneezy 2017) also finds that reports about a state of mind are significantly higher (column 6).

Information about others' behavior: In a few 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 some 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 8 indicate that this leads to a significant reduction in reports.

Subject pool: Student samples report significantly higher 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

[39] This effect could be because subjects report the highest state of all rolls they did, even though they were instructed that only the very first roll counted for the report (Shalvi et al. 2011, Gächter and Schulz 2016). Similarly, the control rolls could provide an excuse or narrative for the subject to report a higher state without feeling too bad about it. Obviously, even if experimenters did not suggest to roll several times, subjects could have rolled several times and report the highest state anyway (or not roll at all and just report whatever they wanted). Perhaps subjects did not have the idea to roll several times. Or the effect is more subtle, i.e., for a valid narrative one needs an external person to suggest the control rolls.

the difference between students and non-students. Students and non-students differ in many respects. We show below that the student effect is partly due to age. In addition, cognitive skills, socio-economic background, current income, etc. could all be part of it.

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

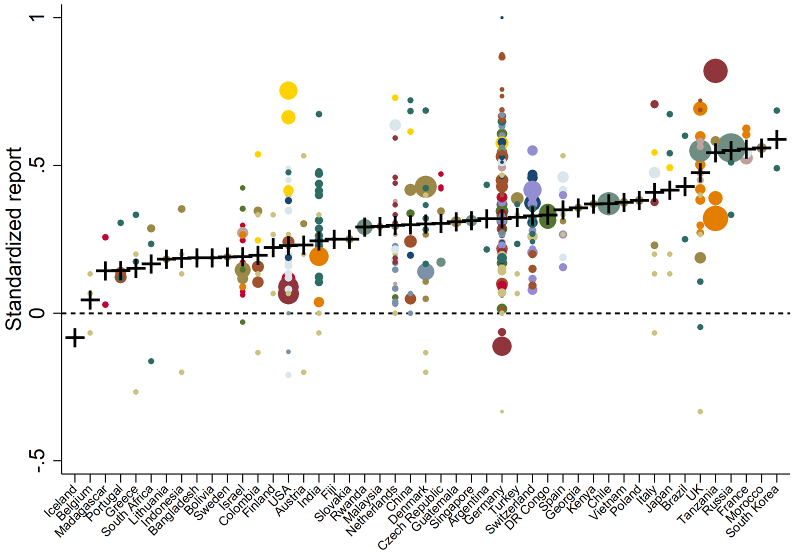
Author affiliation: Studies conducted by economists yield slightly higher reports than studies conducted by psychologists. The differences to sociologists' experiments are not significant.

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.7). Compared to uniform distributions, asymmetric distributions have higher reports and bell-shaped distributions have lower reports.

Country: 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 average standardized report varies only little across countries, from about 0.1 to about 0.5. Adding country fixed effects to the regression in Column 1 of Table A.2 increases the adjusted R^2 from 0.370 to 0.457. For detailed analyses of what drives cross-country differences, see, e.g., Pascual-Ezama et al. (2015), Hugh-Jones (2016), Mann et al. (2016) or Gächter and Schulz (2016).

Figure A.4: Average standardized report by country



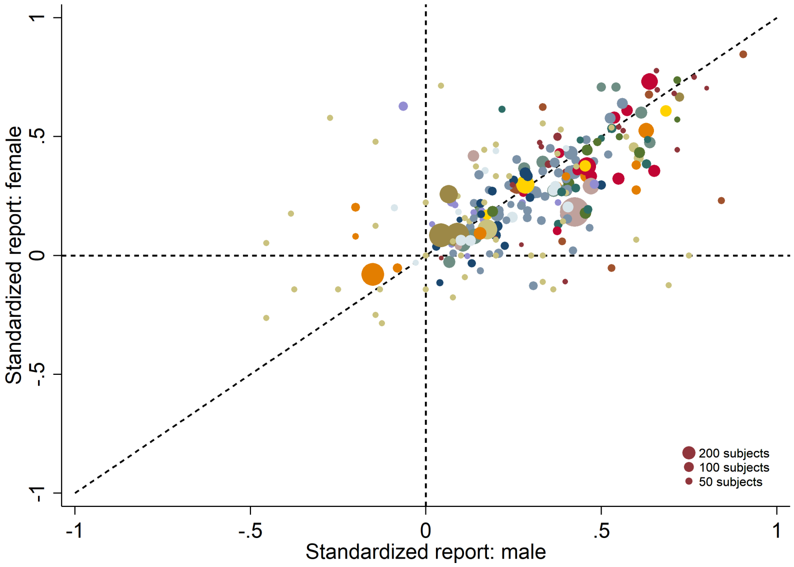
Notes: The figure plots standardized report against country. Standardized report is on the y-axis. A value of 0 means that subjects realize as much payoff as a group of subjects who all tell the truth. A value of 1 means that subjects all report the state that yields the highest payoff. Each bubble represents 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.

A.3 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.

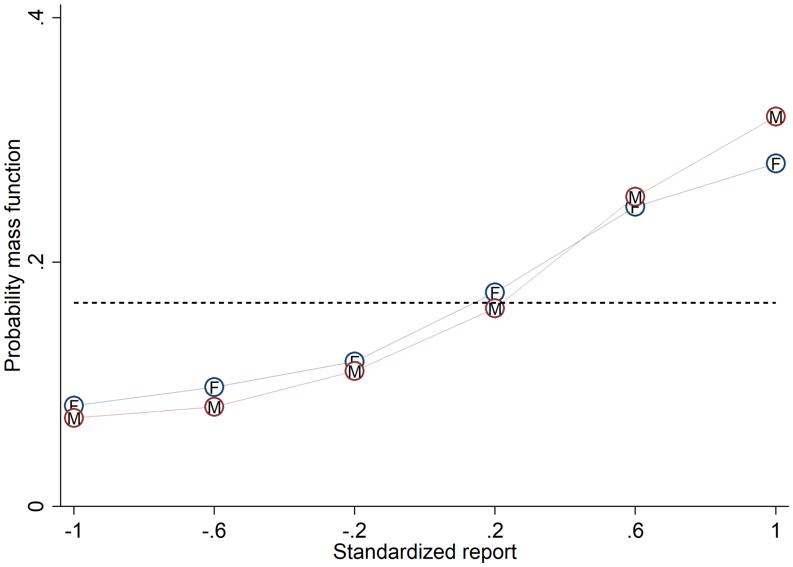
Gender: Figure A.5 shows the effect of gender on reports. The majority of treatments is below the 45° line, indicating that female subjects report lower numbers than male subjects. However, there are also many treatments in which women report higher numbers than men. We test the significance of this 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.6 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.5: 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.6: 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			-0.002	-0.003	
			(0.001)	(0.003)	
Age squared				0.000	
				(0.000)	
1 if economics/management student					0.003
					(0.022)
1 if psychology student					-0.068
					(0.078)
Treatment FE	Yes	Yes	Yes	Yes	Yes
# Decisions	73582	88503	39828	39828	8335
# Subjects	4862	22172	15472	15472	4655
# Treatments	43	239	144	144	52
# Studies	11	47	33	33	9
# Clusters	4806	22116	15472	15472	4655

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.

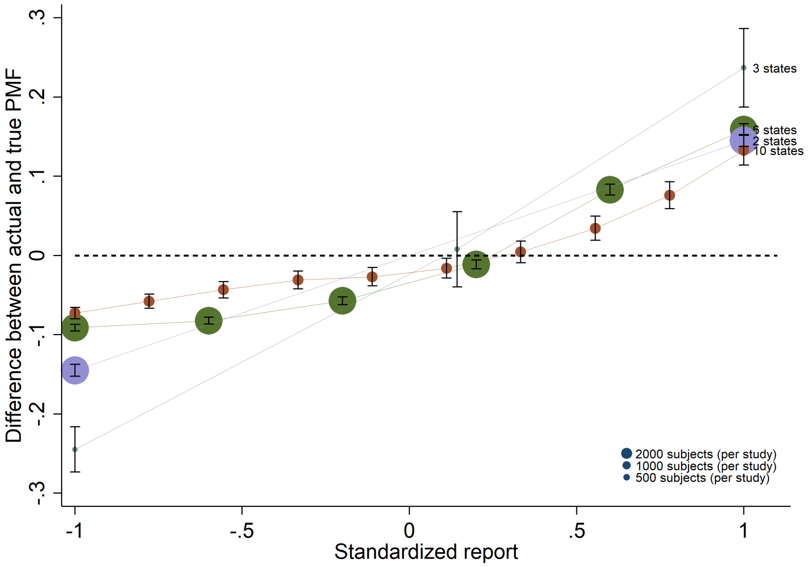
Age: Older subjects tend to report lower numbers. This effect is significant in a linear regression but not significant when we add age squared (Table A.3, columns 3 and 4).

Field of study: While students in general make higher reports than non-students, we do not find an effect of field of study (Table A.3, column 5).

A.4 Further Robustness Checks

Other uniform distributions: In Figure 2, we showed for uniform distributions with two and six states that the distribution of reports is increasing and has support on more than one state (it actually has almost always full support). This finding generalizes to uniform distributions with different number of states. Figure A.7 demonstrates that the distribution of reports is actually quite similar for experiments with different numbers of states. We observe over-reporting of non-maximal states for the six- and 10-state distributions. The general pattern of reporting across the four distributions in the graph suggests that we should expect such over-reporting to occur for any uniform distribution with more than three states.

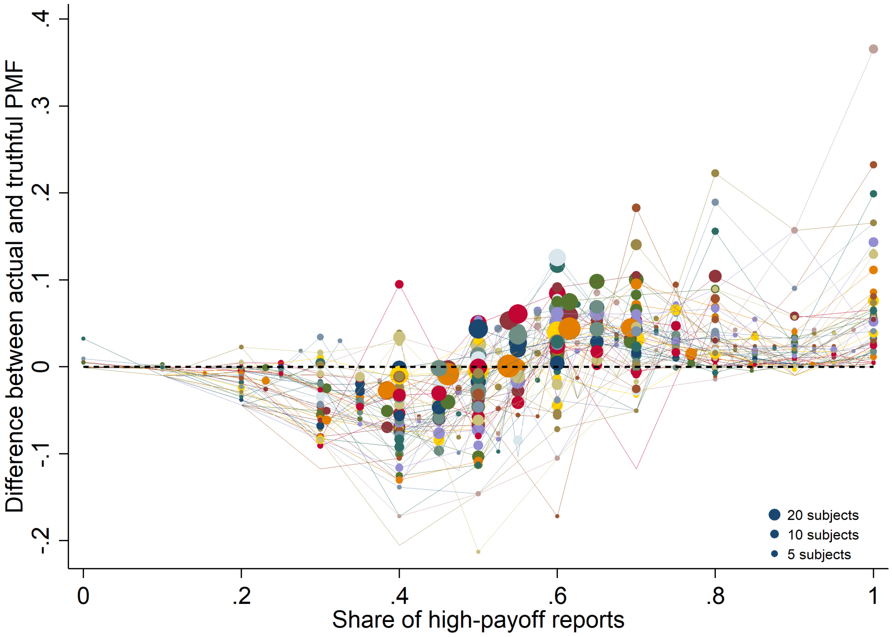
Figure A.7: Distribution of reports (uniform true distributions)



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.

Individual-level analysis: 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 player 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 2-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 the truthful probability mass function. The size of a bubble is proportional to the number of subjects in a given treatment at this share of high-payoff reports.

B Additional Models

In this section we discuss the remaining models listed in Table 2. Proofs are provided immediately after the relevant result. To prove predictions, we first consider binary states, then generalize to n states. Some proofs refer to the proof of Proposition 2 which provides analog results for the LC, the Conformity in LC and the Reputation for Honesty + LC models. Those proofs can be found in Appendix D. Our results also rely on Lemma 1 in Appendix D which states that the results on observability and lying down do not depend on the number of states n.

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 (and so our solution concept is the standard Bayes Nash Equilibrium).[40] Formally, utility is

$$\phi(r, \varsigma(r - \bar{r}); \theta^{IA})$$

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 when $r - \bar{r}$ is 0 and is strictly increasing in the absolute distance from 0 of its argument. The only element of $\vec{\theta}$ that affects utility is the scalar θ^{IA} which governs the weight that an individual applies to inequality aversion. We suppose that ϕ is strictly increasing in its first argument and decreasing in its second (strictly so when $\theta^{IA}>0$), i.e., individuals like money and dislike inequality, and is (weakly) decreasing in θ^{IA}; and that the cross partial of ϕ with respect to the second argument and θ^{IA} is strictly negative, while other cross partials are 0. An equilibrium will exist because of the continuity of ϕ and ς and the fact that $\bar{r}$ is continuous in the distribution of reports.

[40] This model can also capture a notion of a preference for conformity in actions. In this model individuals may gain utility from how closely their action matches others' actions. Because, in this model, an action directly maps to a monetary payoff, caring about the average action of others is the same as caring about the average payoff of others.

Because of the dependence of any given individual's optimal report on others' reports, 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.

Proposition 3 *Suppose individuals have Inequality Aversion utility. For arbitrary n, we have f-invariance, depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance, we have o-invariance and lying down when the state is unobserved or observed. For $n = 2$, the Inequality Aversion model exhibits affinity.*

Proof: We first consider $n = 2$. We will refer to the component of utility coming from inequality aversion as the inequality aversion cost. Observe that utility does not depend directly on the drawn state ω.

Claim 1: Fixing an equilibrium, either all types report r_1, all types report r_2 or there exists one unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.

Consider the case where some individuals give either report. Then by continuity there must be at least one type, $\bar{\theta}^{IA}$, which is indifferent between the two reports. Analogous reasoning to the proof of the LC model-part of Proposition 2 (Appendix D) demonstrates that this type must be unique. By assumption $\frac{\partial^2 \phi}{\partial \varsigma \partial \theta} < 0$ and $\frac{\partial^2 \phi}{\partial r \partial \theta} = 0$. Therefore, since $\phi(r_2, \varsigma(r_2 - \bar{r}); \bar{\theta}^{IA}) - \phi(r_1, \varsigma(r_1 - \bar{r}); \bar{\theta}^{IA}) = 0$, then for all $\theta^{IA} > \bar{\theta}^{IA}$, $\phi(r_2, \varsigma(r_2 - \bar{r}); \bar{\theta}^{IA}) - \phi(r_1, \varsigma(r_1 - \bar{r}); \bar{\theta}^{IA}) < 0$ and for all $\theta^{IA} < \bar{\theta}^{IA}$, $\phi(r_2, \varsigma(r_2 - \bar{r}); \bar{\theta}^{IA}) - \phi(r_1, \varsigma(r_1 - \bar{r}); \bar{\theta}^{IA}) > 0$. Thus the type must be unique.

Claim 2: An equilibrium exists.

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 θ^{IA}). However, the equilibrium may not be unique.

Claim 3: We observe f-invariance.

By Claim 1, the indifferent type (if there is one) must be 0-mass. Since all other individuals have a strict preference, and utility does not depend on the drawn state (and hence does not depend on F), the distribution of reports does not depend on F. Thus the set of equilibria will not change with F.

Claim 4: We observe affinity.

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 $\hat{G}$. ς has a minimum when $r = \bar{r}$. Observe that $r_1 \leq \bar{r} \leq r_2$. Thus, when $\bar{r}$ increases, $|r_1 - \bar{r}|$ increases and $|r_2 - \bar{r}|$ decreases. Thus $\varsigma(r_1 - \bar{r})$ must rise, and $\varsigma(r_2 - \bar{r})$ must fall. Therefore, for all individuals the utility of reporting r_2 increases, and the utility of reporting r_1 decreases, and so more individuals report r_2.

Claim 5: The model exihibits o-invariance and will exhibit downwards lying regardless of observability.

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. Moreover, in any equilibrium with full support on the reporting distribution, we must have some individuals lying down. Since individuals' utility only depends on their report and not their drawn state, generically individuals (other than the zero mass of individuals who are indifferent between reports) with the same parameter θ^{IA} must take the same action. Since we have full support in the reporting distribution, there is some interval of types $[\hat{\theta}^{IA}, \tilde{\theta}^{IA}]$ that strictly prefer to report r_1 over all other reports. Because F features full support, at least some individuals who have $\theta^{IA} \in [\hat{\theta}^{IA}, \tilde{\theta}^{IA}]$ must have drawn $\omega > \omega_1$.

Turning to n states, observe that the reasoning for the f-invariance result is exactly the same (because the set of indifferent types is measure 0, and utility does not depend on the drawn state).

Claim 6: Depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance.

We've already presented an example of affinity for $n = 2$. We now present an example of aversion.

Suppose $n = 3$, and $r_1 = \omega_1 = 0$, $r_2 = \omega_2 = 1$, $r_3 = \omega_3 = 2$. Suppose that utility is equal to $r - \theta^{IA}\varsigma(r - \bar{r})$. We now construct a cost function that is a continuous approximation of the following function: $\varsigma(r - \bar{r}) = 0$ for $|r - \bar{r}| \leq 0.6$, $\varsigma(r - \bar{r}) = 3$ otherwise. Thus, we set $\varsigma(0) = 0$. Then ς increases (in a continuous fashion) so that for a very small δ, when $|r - \bar{r}| = 0.6 - \delta$, $\varsigma(r - \bar{r}) = \epsilon$ (for a very small ϵ). At that point ς increases to 3 at $|r - \bar{r}| = 0.6$, and then ς asymptotes to $3 + \epsilon$ as $|r - \bar{r}| \to \infty$. Moreover, suppose that as a limit case 10% of individuals have $\theta^{IA} = 0.5$, and the rest have $\theta^{IA} = 1$. Suppose $\hat{G^A}$ is such that $\bar{r} = 0.2$. For small enough

ϵ and δ, the former type of individuals reports $r_3 = 2$, the latter type reports $r_1 = 0$ (since reporting $r_1 = 0$ gives a utility of approximately 0, reporting $r_2 = 1$ gives approximately $1 - 3\theta^{IA}$, and reporting $r_3 = 2$ gives approximately $2 - 3\theta^{IA}$). Now if we shift the beliefs about the reporting distribution so that $\hat{G}^B$ induces $\bar{r} = 0.5$, then the former type reports $r_2 = 1$ and the latter type reports $r_2 = 1$ as well (since reporting $r_1 = 0$ gives approximately 0, reporting $r_2 = 1$ gives approximately 1, and reporting $r_3 = 2$ gives approximately $2 - 3\theta^{IA}$). This implies aversion. By continuity, we can also demonstrate $\hat{g}$-invariance. □

B.2 Inequality Aversion + LC

We extend the simple inequality aversion model we developed 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 Navin Kartik 2009). As solution concept we again consider the standard Bayes Nash Equilibrium because utility only depends on the action profile of the individual and the rest of the population. Formally, utility is

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

where $\bar{r}$ is the mean report. The function ς has the same properties as in the Inequality Aversion model. The function c has the same properties as in the LC model. The only elements of $\vec{\theta}$ that affect utility are the scalars θ^{IA} and θ^{LC} which govern the weight that an individual applies to inequality aversion and lying costs. We suppose that ϕ is strictly increasing in its first argument, decreasing in its second (strictly so when $\theta^{IA} > 0$), decreasing in its third (strictly so when $\theta^{LC} > 0$), and is (weakly) decreasing in θ^{IA} and θ^{LC}. Moreover, as before, the partial of ϕ with respect to ς and θ^{IA} is strictly negative and the partial with respect to c and θ^{LC} is strictly negative, while other cross partials are 0. As in the Inequality Aversion model, an equilibrium will exist because of the continuity of ϕ, c and ς and the property that $\bar{r}$ is continuous in the threshold types, but because of the dependence of utility on others' reports, there may be multiple equilibria.

Proposition 4 *Suppose individuals have Inequality Aversion + LC utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we may have affinity, aversion or $\hat{g}$-invariance, we have o-invariance and, depending on parameters, we may* *have lying down or not when the state is unobserved or observed. For $n = 2$, the Inequality Aversion + LC model exhibits drawing in when the equilibrium is unique and affinity.*

Proof: We first consider $n = 2$.

We can define a "threshold function" for each state $\tau_{\omega_i}(\theta^{IA}, \theta^{LC})$, which, given the equilibrium and an individual's given type, gives the utility of reporting $r_{j \neq i}$ versus r_i, conditional on having drawn ω_i. These are continuous functions. If τ is less than or equal to 0, the individual will report their state, otherwise they will lie.

Claim 1: Fixing θ^{IA} and an equilibrium, $\phi(r_2, \varsigma(r_2 - \bar{r}), c(r_2, \omega_1); \theta^{IA}, \theta^{LC}) - \phi(r_1, \varsigma(r_1 - \bar{r}), c(r_1, \omega_1); \theta^{IA}, \theta^{LC})$ is decreasing in θ^{LC}.

The monetary difference between reporting r_1 or r_2 is independent of θ^{LC} as is the inequality aversion cost. But the lying cost part does depend on it: Since $c(r_2, \omega_1) > c(r_1, \omega_1)$, $\frac{\partial \phi}{\partial c} < 0$, $\frac{\partial^2 \phi}{\partial r \partial \theta^{LC}} = 0$ and $\frac{\partial^2 \phi}{\partial c \partial \theta^{LC}} < 0$, the result follows.

The analogous claim holds for those individuals who drew ω_2.

Claim 2: Fixing θ^{LC} and an equilibrium, $\phi(r_2, \varsigma(r_2 - \bar{r}), c(r_2, \omega_1); \theta^{IA}, \theta^{LC}) - \phi(r_1, \varsigma(r_1 - \bar{r}), c(r_1, \omega_1); \theta^{IA}, \theta^{LC})$ is either monotonically increasing or monotonically decreasing in θ^{IA}.

The reasoning is exactly analogous to Claim 1, except that whether we have increasing or decreasing depends on whether $\varsigma(r_1 - \bar{r})$ or $\varsigma(r_2 - \bar{r})$ is larger.

Claim 3: Fixing θ^{LC} and an equilibrium, $\tau_{\omega_i}(\theta^{IA}, \theta^{LC})$ is equal to 0 for at most one value of θ^{IA}. Similarly fixing θ^{IA}, $\tau_{\omega_i}(\theta^{IA}, \theta^{LC})$ is equal to 0 for at most one value of θ^{LC}.

This is immediately implied by the preceding claims.

We can think of the equilibrium as now being characterized by a set of combinations of θ^{LC}s and θ^{IA}s, which conditional on a drawn state imply that decision makers with those parameters are indifferent between the two reports ($\tau_{\omega_i}(\theta^{IA}, \theta^{LC})$=0). We can think of this set as being a function in the space $\theta^{IA} \times \theta^{LC}$; or graphically, a curve in two-dimensional Euclidean space. The LC portion of costs never depends on the distribution of responses, however the rest of the function can.

Because the LC portion of the cost function doesn't depend on the reports of others, we can also think of an equilibrium as the fixed point of the function $\zeta(\bar{r})$ which maps from an aggregate average report to the optimal aggregate average report (given F and H). More precisely, ζ is a function that gives the optimal aggregate average report if there exists one in the allowed range of $\bar{r}$ (i.e. r_1 to r_n); gives r_n if the threshold is above the range; and gives r_1 if the threshold is below the range. This ensures ζ maps from $[r_1, r_n]$ to itself. It also implies that, with a unique equilibrium, the graph of ζ must cross the 45-degree line from above to below.

Claim 4: An equilibrium exists.

An equilibrium will exist because of the continuity of ϕ, c, and ς and the property that $\bar{r}$ is continuous in the threshold types.

Claim 5: Fixing a $\bar{r}$, any individual who draws ω_1 and reports r_2 would also report r_2 if they drew ω_2.

Observe that the utility gap between the two reports if ω_1 is drawn is

$\phi(r_2, \varsigma(r_2 - \bar{r}), c(r_2, \omega_1); \theta^{IA}, \theta^{LC}) - \phi(r_1, \varsigma(r_1 - \bar{r}), c(r_1, \omega_1); \theta^{IA}, \theta^{LC})$. The gap if ω_2 is drawn is $\phi(r_2, \varsigma(r_2 - \bar{r}), c(r_2, \omega_2); \theta^{IA}, \theta^{LC}) - \phi(r_1, \varsigma(r_1 - \bar{r}), c(r_1, \omega_2); \theta^{IA}, \theta^{LC})$. By construction the latter utility gap is larger than the former.

Claim 6: We observe drawing in.

Suppose the equilibrium is unique and that $f(\omega_2)$ increases while fixing strategies. Consider what happens to $\zeta(\bar{r})$. There are more individuals drawing the high state, and fewer drawing the low state. Since individuals are more likely to report high after having drawn the high state than the low state by Claim 5 (since the set of individuals who draw ω_2 and report r_2 is a superset of those who would report r_2 if they drew ω_1), this implies an increase in the optimal aggregate report $\bar{r}$ (i.e., $\zeta(\bar{r})$). This implies that $\varsigma(r_2 - \bar{r})$ gets smaller and $\varsigma(r_1 - \bar{r})$ gets larger, and so r_2 becomes relatively more attractive to all individuals. This also increases $\zeta(\bar{r})$. Thus, the increase in $f(\omega_2)$ shifts ζ up and so the equilibrium level of $\bar{r}$ increases. Whenever $\bar{r}$ increases, reporting r_2 becomes relatively more attractive (since $\varsigma(r_1 - \bar{r})$ increases and $\varsigma(r_2 - \bar{r})$ falls, causing drawing in.

Although the equilibrium may not be unique, because G enters in the utility function directly (through its one-to-one mapping with $\bar{r}$) we can still make predictions regarding the effect of $\hat{G}$.

Claim 7: We observe affinity.

If $\hat{g}(r_2)$ increases, then the beliefs about $\bar{r}$ increases. This implies that $\varsigma(r_2 - \bar{r})$ gets smaller and $\varsigma(r_1 - \bar{r})$ gets larger, and so r_2 becomes relatively more attractive to all individuals.

Claim 8: The model exihibits o-invariance and may exhibit downwards lying or not re-

gardless of observability.

Because this model nests the standard Inequality Aversion model, if individuals dislike being too far ahead of others, they may lie down. But the model also nests the LC model where individuals will never lie down. Moreover, as in the LC model and the Inequality Aversion model individually, the distribution of reports does not depend on observability.

Now we turn to n states.

Claim 9: Depending on parameters, we may have drawing in, drawing out or f-invariance.

For $n = 2$, we have shown that drawing in will occur. We now provide an example for $n = 3$ that yields drawing out. Consider the limiting case where the vast majority of individuals have just LC utility and some individuals have utility that takes into account only inequality aversion costs, where the inequality aversion cost is a function of the absolute distance between an individual's report and the average report. Moreover, suppose the parameters of the LC costs (for all individuals) are such that individuals who care only about LC costs are willing to lie up two states, but no one is willing to lie up one state (e.g., because of fixed costs). In contrast, we suppose that the inequality aversion costs are large enough so that those individuals simply want to match as closely as possible the average report.

Thus, all individuals with only LC costs who drew ω_1 will report r_3 regardless of what others do. Individuals with only LC costs who drew ω_2 (ω_3 respectively) will report r_2 (r_3 respectively). Suppose that we have a distribution where $f(\omega_1)$ is close to 1; then $\bar{r}$ is closer to r_3 than r_2 regardless of what the inequality averse individuals do, and so those individuals face a relatively strong incentive to lie up all the way to the highest state r_3. Now suppose we shift much of the weight of F from ω_1 to ω_2. Now those individuals facing only LC costs who previously drew ω_1 but now draw ω_2 will report r_2 instead of r_3. This can shift $\bar{r}$ closer to r_2 than r_3 (regardless of what the inequality averse individuals do), and so now inequality averse individuals will report r_2. Thus, we get drawing out. By continuity, we can also have f-invariance.

Claim 10: Depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance.

Since this model nests the Inequality Aversion model as a limit case, and since that model can generate affinity, aversion or $\hat{g}$-invariance, this model can too. □

B.3 Censored Conformity in LC

This section presents a variation of the Conformity in LC model. One could imagine that an individual does not normalize their lying cost by the average lying cost in society (as in Conformity in LC), but only by the lying costs incurred by individuals who "could have" lied profitably, i.e., those who did not receive the maximal draw. As in the Conformity in LC model, utilities thus depend on the profile of joint state-report combinations across other individuals, and so we solve for the Bayes Nash Equilibrium.

In this model, as in Conformity in LC, individuals will not want to lie downwards. We denote, supressing extraneous notation, the average lying costs of all those who do not draw the maximal state as $\bar{c}_{\omega \neq \omega_n}$. The utility function is then:

$$\phi(r, \eta(c(r, \omega), \bar{c}_{\omega \neq \omega_n}); \theta^{CCLC})$$

where η is the normalized cost function and has the same properties as in the Conformity in LC model. ϕ is strictly increasing in the first argument, falling in the second (strictly when $\theta^{CCLC} > 0$), and (weakly) falling in θ^{CCLC}. Last, the cross partial of ϕ with respect to η and θ^{CCLC} is strictly negative, while other cross partials are 0. An equilibrium will exist because of the continuity of ϕ, η and c (and the continuity of $\bar{c}_{\omega \neq \omega_n}$ in the proportion of liars), but because of the dependence of utility on others' joint state-report combinations, it may not be unique.

Proposition 5 *Suppose individuals have Censored Conformity in LC utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we may have affinity, aversion or $\hat{g}$-invariance, we have o-invariance and no lying down when the state is unobserved or observed. For $n = 2$, we have f-invariance and affinity.*

Proof: We first consider $n = 2$.

Claim 1: No individual lies down.

In doing so they would pay a weakly higher lying cost and receive a lower monetary payoff than if they told the truth.

Claim 2: Fixing an equilibrium, conditional on drawing ω_1 either all types report r_1, all types report r_2 or there exists one unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.

Observe that since no one lies down, the fraction of individuals who lie among those who could lie is simply the excess number of r_2 reports compared to ω_2 draws, divided by the number of individuals drawing ω_1: $\frac{g(r_2)-f(\omega_2)}{f(\omega_1)}$. The actual lying cost, conditional on those that could have lied, is proportional to this (for $n = 2$), a proportionality we can directly model as part of ϕ. In the case that some types give one report and others the other, by continuity there must be a type that conditional on drawing ω_1 is indifferent between the two reports. This type $\bar{\theta}^{CCLC}$ satisfies:

$$\phi(r_2, \eta(c(r_2, \omega_1), \frac{g(r_2) - f(\omega_2)}{f(\omega_1)}); \bar{\theta}^{CCLC}) = \phi(r_1, \eta(c(r_1, \omega_1), \frac{g(r_2) - f(\omega_2)}{f(\omega_1)}); \bar{\theta}^{CCLC})$$

This threshold is unique for the analogous reasons to the LC and Conformity in LC models.

We can rewrite the indifference condition as

$$\phi(r_2, \eta(c(r_2, \omega_1), H(\bar{\theta}^{CCLC})); \bar{\theta}^{CCLC}) = \phi(r_1, \eta(c(r_1, \omega_1), H(\bar{\theta}^{CCLC})); \bar{\theta}^{CCLC})$$

By construction $H(\bar{\theta}^{CCLC})$ is the fraction of subjects who would report r_2 if they drew ω_1. And so we have $H(\bar{\theta}^{CCLC}) = Prob(\theta < \bar{\theta}^{CCLC}) = \frac{f(\omega_1)}{f(\omega_1)} Prob(\theta < \bar{\theta}^{CCLC}) = \frac{g(r_2)-f(\omega_2)}{f(\omega_1)}$.

Claim 3: An equilibrium exists.

An equilibrium will exist given the continuity of ϕ and η and the property that the proportion of liars is continuous in the cutoff $\bar{\theta}^{CCLC}$

Claim 4: The model exhibits f-invariance.

The indifference condition in Claim 2 does not depend on F and we obtain f-invariance.

Claim 5: The model exhibits affinity.

As in the standard Conformity in LC model, the equilibrium reporting distribution may not be unique. We can still make predictions regarding the effect of $\hat{G}$ since no one lies down. Suppose we fix F and $\hat{g}(r_2)$ increases. Then there must be more liars who drew ω_1 and said r_2 and so the second argument of the utility function must increase. Thus, the cost of lying goes down. Previously indifferent type must strictly prefer to lie, which yields affinity.

Claim 6: The model exihibits o-invariance and no downwards lying regardless of observability.

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 Conformity in LC model.

We now turn to $n > 2$.

Claim 7: Depending on parameters, we may have drawing in, drawing out or f-invariance.

Observe that the example of drawing in provided in Claim 12 of the Conformity in LC model proof (Proposition 2 in Appendix D) relied on the aggregate lying costs going up for those individuals who could lie. This implies that it works just as well in this model. We could reverse the example to obtain drawing in. By continuity, we can also generate f-invariance.

Claim 8: Depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance.

We demonstrated affinity already. The example for aversion provided in Claim 13 of the Conformity in LC model works here as well. By continuity, we can also generate $\hat{g}$-invariance. □

B.4 Reputation for Being Not Greedy

Individuals often want to signal to the audience about a particular characteristic they possess. We use as an inspiration the motivations provided in Bénabou and Tirole (2006) and Fischbacher and Föllmi-Heusi (2013), and model an individual as wanting to signal to the audience that they are not greedy, i.e., they place a relatively low value on money compared to reputation. Thus, an individual's utility will depend on the audience's beliefs about their type, the scalar θ^{RNG} (the only element of $\vec{\theta}$ that affects utility), which is unobserved by the audience. However, the belief can be conditioned on the report r itself. Because utility depends on the audience's beliefs, we must use the psychological game theory framework of Battigalli and M. Dufwenberg (2009) to analyze the game. Since the audience player understands the equilibrium strategies of all types, and correctly utilizes Bayesian updating, we can simply describe their belief as $E(\theta^{RNG}|r)$. Given this, utility is:

$$\phi(r, E(\theta^{RNG}|r); \theta^{RNG})$$

We assume ϕ is increasing in the first element, i.e., individuals like money; but the partial of ϕ with respect to the first element is equal to 0 when $\theta^{RNG} = \kappa^{RNG}$, and otherwise strictly positive for $\theta^{RNG} < \kappa^{RNG}$. ϕ is also increasing in the second element, i.e., individuals like the audience to have a high belief about their θ^{RNG}; specifically the partial of ϕ with respect to the second element is 0 when $\theta^{RNG} = 0$ and is strictly positive for $\theta^{RNG} > 0$. The cross partial of ϕ with respect to the first element and θ^{RNG} is strictly negative. This captures the property that individuals face both a higher benefit, and a higher marginal benefit, of the monetary payoff when θ^{RNG} is smaller. Moreover, the cross partial of ϕ with respect to the second element and θ^{RNG} is strictly positive. This captures the property that individuals with higher θ^{RNG}s have both a higher benefit, and a higher marginal benefit, of being perceived as having a higher expected θ^{RNG}. Other cross partials are 0. Intuitively our assumptions are tantamount to supposing that less "greedy" individuals also care more about being thought of as less greedy. Equilibrium will exist because of the continuity of ϕ and the expectations operator, but may not be unique.

Proposition 6 *Suppose individuals have Reputation for Being Not Greedy utility. For arbitrary n, we have f-invariance, depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance, we have o-invariance and lying down when the state is unobserved or observed.*

Proof: We first consider $n = 2$.

Claim 1: Fixing an equilibrium, either all types report r_1, all types report r_2 or there exists one unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.

Consider the case where some individuals give either report. Then by continuity there must be at least one type, $\bar{\theta}^{RNG}$, which is indifferent between the two reports: $\phi(r_2, E(\theta^{RNG}|r_2); \bar{\theta}^{RNG}) - \phi(r_1, E(\theta^{RNG}|r_1); \bar{\theta}^{RNG}) = 0$. With full support on G this implies that $E(\theta^{RNG}|r_2) < E(\theta^{RNG}|r_1)$. If not, then reporting r_2 gives higher utility and so all types give report r_2, a contradiction. Then for all $\theta^{RNG} > \bar{\theta}^{RNG}$, $\phi(r_2, E(\theta^{RNG}|r_2); \theta^{RNG}) - \phi(r_1, E(\theta^{RNG}|r_1); \theta^{RNG}) < 0$ and for all $\theta^{RNG} < \bar{\theta}^{RNG}$, $\phi(r_2, E(\theta^{RNG}|r_2); \theta^{RNG}) - \phi(r_1, E(\theta^{RNG}|r_1); \theta^{RNG}) > 0$ by our assumptions on the cross partials.

Claim 2: An equilibrium exists.

This is by standard continuity arguments.

Claim 3: We have f-invariance.

Observe that utility does not depend directly on the drawn state ω. With reasoning analogous to that given in the Inequality Aversion model the reporting strategy thus also does not depend on ω for all but a 0-mass of individuals (those who are indifferent). Even though there can be multiple equilibria, this implies that the distribution of reports does not depend on F and so the set of equilibria will not change with F.

Claim 4: Depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance.

The uniqueness of the equilibrium depends on H — because the construction of the indifferent type depends on the relationship between the expectation of θ^{RNG}, conditional on it being above the indifferent type, and the expectation of θ^{RNG}, conditional on it being below the indifferent type. The actual shape of H can be such that there are multiple equilibria or a unique equilibrium. Importantly though, recall that conditional on a particular equilibrium there is one unique indifferent type. The reason why we may get affinity, aversion or $\hat{g}$-invariance is that a shift in $\hat{G}$ could be rationalized by different shifts in H that could lead to either affinity or aversion. We provide an example.

Suppose the support for θ^{RNG} is $[0, 1]$, that utility from report r is $\theta^{RNG}E[\theta^{RNG}|r] + (1 - \theta^{RNG})r$ and there are binary states/reports (with payoffs of $r_2 = 1$ and $r_1 = 0$). As claimed above, it is easy to verify that, in any equilibrium with full support, there is a single unique indifferent type that satisfies $\theta^{RNG}E[\theta^{RNG}|0] = \theta^{RNG}E[\theta^{RNG}|1] + (1 - \theta^{RNG})$ or $\theta^{RNG}(1 + E[\theta^{RNG}|0] - E[\theta^{RNG}|1]) = 1$. Moreover, also as claimed above, $E[\theta^{RNG}|0] - E[\theta^{RNG}|1] > 0$ in any equilibrium with full support.

Now, we show that we can either have affinity or aversion. Suppose that $\hat{g}(r_2)$ increases. This implies there is a larger mass of individuals below the threshold than previously. This could be rationalized by different shifts in H which induce different reactions. For example, individuals could be less likely to draw a value just above the threshold, and more likely to draw values far below the threshold. This implies that the value of θ^{RNG}, conditional on reporting $r_1 = 0$, has gone up, and the value of θ^{RNG}, conditional on reporting $r_2 = 1$, has gone down, implying that $1 + E[\theta^{RNG}|0] - E[\theta^{RNG}|1]$ has increased. Thus, the indifferent type must fall.

However, another way to rationalize the shift in behavior is there are fewer individuals with very high types (θ^{RNG} close to 1), and many more individuals with types just below the threshold. This implies that the value of θ^{RNG}, conditional on reporting $r_1 = 0$, has gone down, and the value of θ^{RNG}, conditional on reporting $r_2 = 1$, has gone up, implying that

$1 + E[\theta^{RNG}|0] - E[\theta^{RNG}|1]$ has decreased, thus the indifferent type must increase.[41] Thus, observing a higher $\hat{g}(r_2)$ could either increase or decrease the threshold. By continuity, we can also generate $\hat{g}$-invariance.

Claim 5: The model exihibits o-invariance and will exhibit downwards lying regardless of observability.

Some individuals will lie downwards in an equilibrium with full support since if a given type (other than the indifferent type, which has 0 mass) prefers to report r_1, conditional on drawing ω_1, the same type would want to report r_1, conditional on drawing ω_2. 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 the audience player with inferring the type of the individual, and again not change the set of possible equilibria and the predictions regarding downward lying is the same under observability.

The previous predictions do not depend on the number of states, so they also apply for arbitrary n states. □

B.5 LC-Reputation

Rather than caring about the reputation of having reported truthfully conditional on their report, individuals may instead want to cultivate a reputation as a person who has high lying costs, i.e., they like the audience to have a high belief about their θ^{LC}. Such a model is similar to the one discussed in Frankel and Narvin Kartik (forthcoming). 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. Thus, an individual's utility will depend on the audience player's beliefs about their lying cost type, the scalar θ^{LC}, which is unobserved. However, the belief can be conditioned on the report r itself. Because utility depends on the audience's beliefs, we use the psychological game theory framework of Battigalli and M. Dufwenberg (2009) to analyze the game. Since the audience understands the equilibrium strategies of all types, and correctly utilizes Bayesian updating, we can simply describe their

[41] For similar reasons, $1 + E[\theta^{RNG}|0] - E[\theta^{RNG}|1]$ may be non-monotone in the threshold type.

belief as $E(\theta^{LC}|r)$.

Utility is

$$\phi(r, c(r,\omega), E[\theta^{LC}|r]; \theta^{LC}, \theta^{Rep}) = u(r) - \theta^{LC} c(r,\omega) + \theta^{Rep} \upsilon(E[\theta^{LC}|r])$$

The only elements of $\vec{\theta}$ that affect utility are θ^{LC} and θ^{Rep}. $u(r)$ is strictly inreasing in r. c and θ^{LC} have the same interpretation as in the LC model, and the assumptions regarding them are the same. θ^{Rep} represents the weight that any given individual places on the audience's belief about θ^{LC}. υ is strictly increasing in its argument. The interpretation is that individuals have a positive utility from others believing that they have high lying costs. An equilibrium will exist because of the continuity of ϕ, c and the expectations operator, but may not be unique because of the dependence of utility on others' strategies (via the audience's beliefs).

Proposition 7 *Suppose individuals have LC-Reputation utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we may have affinity, aversion or $\hat{g}$-invariance, we have o-shift and, depending on parameters, we may have lying down or not when the state is unobserved or observed. For $n = 2$, the LC-Reputation model predicts drawing in.*

Proof: We first consider $n = 2$.

Claim 1: $E[\theta^{LC}|r_1] \geq E[\theta^{LC}|r_2]$ for all equilibria with full support.

To see this, suppose not. Then r_2 has both a (strictly) higher reputation and (strictly) higher monetary payoff. Fix a value of θ^{Rep}. All those who drew ω_2 will report r_2. Observe that by reasoning analogous to the LC model itself, fixing θ^{Rep} and an equilibrium, $\phi(r_2, c(r_2,\omega_1), E[\theta^{LC}|r_2]; \theta^{LC}, \theta^{Rep}) - \phi(r_1, c(r_1,\omega_1), E[\theta^{LC}|r_1]; \theta^{LC}, \theta^{Rep})$ is decreasing in θ^{LC}. Thus, of those who drew ω_1, there will be a threshold type and all types with a higher θ^{LC} will report r_1, all those with a lower type will report r_2. But this immediately implies that $E[\theta^{LC}|r_1, \theta^{Rep}] \geq E[\theta^{LC}|r_2, \theta^{Rep}]$ and so, averaging over values of θ^{Rep}, we obtain $E[\theta^{LC}|r_1] \geq E[\theta^{LC}|r_2]$.

Claim 2: Fixing θ^{LC} and an equilibrium, $\phi(r_2, c(r_2,\omega_1), E[\theta^{LC}|r_2]; \theta^{LC}, \theta^{Rep}) - \phi(r_1, c(r_1,\omega_1), E[\theta^{LC}|r_1]; \theta^{LC},$ is decreasing in θ^{Rep}.

This is immediately implied by the fact that the reputation is worse at r_2 (as shown in Claim 1), $\frac{\partial \phi}{\partial E[\theta^{LC}]} > 0$, $\frac{\partial^2 \phi}{\partial E[\theta^{LC}]\partial \theta^{Rep}} > 0$ and the other cross partials with respect to θ^{Rep} are 0 (by our assumption of additive separability).

As in the Reputation for Honesty + LC model (see proof of Proposition 2 in Appendix D), we can construct a "threshold function" for each state $\tau_{\omega_i}(\theta^{LC}, \theta^{Rep})$ which, given the equilibrium and an individual's type, gives the utility of reporting $r_{j \neq i}$ versus r_i, conditional on having drawn ω_i.

Claim 3: Fixing θ^{LC} and an equilibrium, $\tau_{\omega_i}(\theta^{LC}, \theta^{Rep})$ is equal to 0 for at most one value of θ^{Rep}. Similarly fixing θ^{Rep}, $\tau_{\omega_i}(\theta^{LC}, \theta^{Rep})$ is equal to 0 for at most one value of θ^{LC}.

This is immediately implied by the preceding claims.

If τ is less than or equal to 0, 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^{LC}s and θ^{Rep}s so that the threshold function equals 0. Thus the threshold diagram looks qualitatively similar to Figure D.1 (including the linear threshold functions).

We can characterize the equilibrium in terms of the intercepts of the threshold function. Observe that given H and a utility function, $E[\theta^{LC}|r_i]$ is characterized by the function $\tau_{\omega_i}(\theta^{LC}, \theta^{Rep}) = 0$. Since the $\tau_{\omega_i}(\theta^{LC}, \theta^{Rep}) = 0$ equations are always linear in θ^{LC} and θ^{Rep} they can be characterized by its θ^{LC} intercept and its θ^{Rep} intercept denoted $\theta^{\omega_i}_{LC,T}$ and $\theta^{\omega_i}_{Rep,T}$. Moreover, since the LC portion of costs never depends on the distribution of responses, the $\theta^{\omega_i}_{LC,T}$ intercept (i.e. the threshold value of $\theta^{\omega_i}_{LC,T}$ when $\theta^{Rep} = 0$) must always be the same. Therefore, we can think of each of the threshold "lines" (one for each drawn state) as being characterized by a single intercept: $\theta^{\omega_i}_{Rep,T}$. The thresholds $\theta^{\omega_i}_{Rep,T}$ (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 estimated value of θ^{LC} at either report (as well as G).

To solve for an equilibrium we can consider a function $\zeta(\theta^{\omega_1}_{Rep,T}, \theta^{\omega_2}_{Rep,T})$ which maps from the thresholds that everyone is using into best response thresholds. The fixed points of this function will characterize our equilibria. However, observe that because we are looking at the θ^{Rep} intercepts, the LC 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 $\theta^{\omega_1}_{Rep,T} = \theta^{\omega_2}_{Rep,T}$. Thus, our problem reduces to a single dimension; and we can consider a function $\zeta(\theta_{Rep,T})$, and its fixed points characterize the equilibria. Thus, ζ is a function that gives the optimal threshold if there exists one in the allowed range of θ^{Rep}; gives κ^{Rep} if the threshold is above the range; and gives 0 if the threshold is below the range. This ensures ζ maps from $[0, \kappa^{Rep}]$ to itself. Moreover, if there is a unique equilibrium, the graph of ζ must cross the 45-degree line from above to below.

Claim 4: An equilibrium exists.

Given our continuity assumptions, the threshold functions will be continuous in the conditional expectations of θ^{LC}, and the conditional expectations will be continuous in the threshold functions, so an equilibrium will exist. However, the equilibrium may not necessarily be unique.

Claim 5: We observe drawing in.

Suppose there is a unique equilibrium. Recall that $E[\theta^{LC}|r_1] \geq E[\theta^{LC}|r_2]$. Moreover, observe that fixing θ^{Rep}, $E[\theta^{LC}|r_2, \omega_2, \theta^{Rep}] \geq E[\theta^{LC}|r_2, \omega_{1,} \theta^{Rep}]$, since only those with low θ^{LC} will lie from ω_1 to r_2. Thus the following is true averaging over θ^{Rep}: $E[\theta^{LC}|r_2, \omega_2] \geq E[\theta^{LC}|r_2, \omega_1]$. Analogous reasoning leads to $E[\theta^{LC}|r_1, \omega_1] \geq E[\theta^{LC}|r_1, \omega_2]$. Now suppose that $f(\omega_2)$ increases. Fixing the input threshold $\theta_{Rep,T}$, this implies that the fraction of individuals, conditional on reporting r_2, who drew ω_2, must increase. Similarly, the fraction of individuals, conditional on reporting r_1, who drew ω_2, must increase. This increases the expected θ^{LC} at r_2 and decreases it at r_1. This makes r_2 relatively more attractive to individuals (compared to r_1). Thus the optimal threshold θ^{Rep} (generated by ζ) must rise and we get drawing in.

Claim 6: Depending on parameters, we may observe affinity, aversion or $\hat{g}$-invariance.

Because the threshold characteristics look qualitatively similar to Figure D.1 we can again see how a shift in $\hat{g}(r_2)$ can cause either affinity, aversion or $\hat{g}$-invariance even when the equilibrium reporting distribution is unique. Consider the threshold $\theta_{Rep,T}$. It is defined as the solution to the equation $u(r_2)+\theta^{Rep}v(E[\theta^{LC}|r_2]) = u(r_1)+\theta^{Rep}v(E[\theta^{LC}|r_1])$ or $u(r_2)-u(r_1) = \theta^{Rep}(v(E[\theta^{LC}|r_1]) - v(E[\theta^{LC}|r_2]))$.

The $\hat{G}$ treatments do not pin down the new belief about H that subjects hold. Depending on the H, we could get affinity or aversion. In particular, suppose we move from $\hat{G}^A$ (associated with H^A) to $\hat{G}^B$ and that there are two Hs (H^B and $\tilde{H}^B$) that rationalize $\hat{G}^B$. It can be the case that under H^B the value $v(E[\theta^{LC}|r_1]) - v(E[\theta^{LC}|r_2])$ is larger than under H^A. In contrast, under $\tilde{H}^B$ the difference is smaller than under H^A. Then we get aversion if subjects believe the new H is the former, and aversion if the latter.

Formally, we show that two different changes in the exogenous distribution H can both lead to an increase in $\hat{g}^B(r_2)$ (relative to $\hat{g}^A(r_2)$). Then we show that they have the opposite implications for $\upsilon(E[\theta^{LC}|r_1]) - \upsilon(E[\theta^{LC}|r_2])$. As in the Reputation for Honesty + LC model two different shifts of probability mass in H could lead to an increase in $\hat{g}^B(r_2)$ (relative to $\hat{g}^A(r_2)$). The first shifts mass from above $\tau(\omega_1)$ to below it (without altering the relative weights above and below $\tau(\omega_2)$). This, fixing the thresholds, doesn't change the reporting of individuals who drew ω_2, but leads to a higher mass of individuals drawing ω_1 reporting r_2. Since $E[\theta^{LC}|r_2, \omega_2] \geq E[\theta^{LC}|r_2, \omega_1]$ and $E[\theta^{LC}|r_1, \omega_1] \geq E[\theta^{LC}|r_1, \omega_2]$ this decreases both $E[\theta^{LC}|r_2]$ and $E[\theta^{LC}|r_1]$, as well as increasing $g(r_2)$. Recall our fixed point operator that defines the threshold which characterizes the equilibrium: $\zeta(\theta_{Rep,T})$. Recall that this, taking as an input everyone else's threshold, returns the optimal threshold. If $\upsilon(E[\theta^{LC}|r_1]) - \upsilon(E[\theta^{LC}|r_2])$ increases, this makes the high report less attractive, and so ζ decreases, reducing the equilibrium level of $\theta_{Rep,T}$.[42] This reduction will cause aversion. Thus, in order to generate aversion we need that $\upsilon(E[\theta^{LC}|r_1]) - \upsilon(E[\theta^{LC}|r_2])$ increases in response to this shift in weight, and as in the Reputation for Honesty + LC model a simple restriction on the derivative of υ at $E[\theta^{LC}|r_1]$ and $E[\theta^{LC}|r_2]$ will suffice.

The second shift moves mass from below $\tau(\omega_2)$ to above it (without altering the relative weights above and below $\tau(\omega_1)$). Fixing the thresholds, this doesn't change the reporting of individuals who drew ω_1, but leads to a higher mass of individuals drawing ω_2 reporting r_2. This increases the expected value of θ^{LC} at both reports. If $\upsilon(E[\theta^{LC}|r_1]) - \upsilon(E[\theta^{LC}|r_2])$ decreases, this makes the high report more attractive, and so ζ increases. This increases the equilibrium level of $\theta_{Rep,T}$, and causes affinity. Similarly to before, in order to generate affinity we need that $\upsilon(E[\theta^{LC}|r_1]) - \upsilon(E[\theta^{LC}|r_2])$ decreases in response to this shift in weight. This again occurs with a simple restriction on the derivative of υ, as in the Reputation for Honesty + LC. Thus, we can get both affinity and aversion (and by continuity $\hat{g}$-invariance).

Claim 7: The model exhibits o-shift and can exhibit downwards lying or not regardless of observability.

Individuals' behavior should change if the state is observed. But this is for a very different reason compared to the Reputation for Honesty + LC model. In that model, behavior changes

[42] An equilibrium threshold must fall in this situation (see the Reputation for Honesty + LC model for details of why).

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 θ^{LC}, as it can be estimated using both ω and r, rather than just r.

Given the similarity to the Reputation for Honesty + 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 not done on the probability of being a liar, as in the Reputation for Honesty + LC model, but on θ^{LC}. It is possible to have a countersignalling equilibrium where the highest and lowest θ^{LC} types pool on truth-telling and middle θ^{LC} types lie down. Of course, if individuals care vary little about their reputation, then we will never observe lying down.

We now turn to n states.

Claim 8: Depending on parameters, we may have drawing in, drawing out or f-invariance.

We have shown drawing in for $n = 2$. We now provide an example for drawing out analogous to that for the Reputation for Honesty + LC model. Suppose that $n = 3$. Moreover, suppose that the LC part of the utility function is such that individuals only lie one state/report up. Now, move from F^A to F^B by keeping $f^A(\omega_1)$ constant and shifting weight from ω_2 to ω_3. This has two effects. First, fixing strategies, it makes reporting r_3 more attractive (since some of the individuals drawing ω_3 will still report r_3) and so increases the estimated value of θ^{LC} at r_3. Second, by the same reasoning, it makes the middle state less attractive. Thus, individuals who draw the lowest state will find reporting the middle state less attractive, and more will simply report the truth which implies drawing out. By continuity, the model can also generate f-invariance.

We know we have ambiguous predictions regarding shifts in $\hat{G}$ for even two states, and this carries over to n states. □

B.6 Guilt Aversion

Guilt aversion (Charness and M. Dufwenberg 2006; Battigalli and M. Dufwenberg 2007, 2009) posits that people like to live up to others' expectations so as to avoid guilt. In applying guilt aversion to our setting, we assume that subjects experience guilt (and so lower utility) to the extent that they believe they disappointed the audience player (i.e., report more than expected), for example, the experimenter. Because beliefs are correct in equilibrium, the audience expects the report to be the average report induced by the equilibrium G, which we denote $\bar{r}$ (each equilibrium will have an associated $\bar{r}$). To keep notation simple, we suppress the fact that $\bar{r}$ is an equilibrium object that depends on F and H and the selected equilibrium.[43] Because individuals' utility depends on the beliefs of the audience, this model explicitly uses the tools of psychological game theory (Battigalli and M. Dufwenberg 2007, 2009). Utility is:

$$\phi(r, \gamma(r - \bar{r}); \theta^{GA})$$

where γ is a function that maps the difference between any given individual's report and the average report to a utility cost. Given an equilibrium and associated $\bar{r}$, if $r \leq \bar{r}$, then $\gamma(r - \bar{r}) = 0$. If $r > \bar{r}$, then $\gamma(r - \bar{r})$ is strictly increasing in $r - \bar{r}$. The only element of $\vec{\theta}$ that affects utility is the scalar θ^{GA} which governs the weight that an individual applies to guilt. We suppose that ϕ is strictly increasing in its first argument, decreasing in its second (strictly so when $\theta^{GA}>0$), (weakly) decreasing in θ^{GA}, and the cross partial of the second argument and θ^{GA} is strictly negative, while other cross partials are 0.

Equilibrium existence follows from the continuity of ϕ and γ and $\bar{r}$. However, there may be multiple equilibria. For example, if the audience expects that the only report given is the maximal report, then players do not believe that the audience will be disappointed when the maximal report is made. Thus no one feels guilt when making the maximal report, and so everyone makes that report. This forms an equilibrium. In contrast, if the audience expects that the only report given is the minimal report, then the audience will be disappointed when any other report is made. So long as individuals experience enough guilt, it can also be an equilibrium for everyone to then make the minimal report. However, as we formalize below, the set of equilibria doesn't shift with F.[44]

[43] One might argue that guilt aversion is not appropriate for this subject-experimenter interaction (or more generally, subject-audience 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 almost a dozen 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 not very far away from 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.

[44] Surprisingly, in our simple environment with our particular modeling assumptions, guilt aversion turns out to predict the same as the inequality aversion model, albeit for very different underlying reasons. The assumption about utilities when $r \leq \bar{r}$ is different but this does not affect the predictions.

Proposition 8 *Suppose individuals have Guilt Aversion utility. For arbitrary n, we have f-invariance, depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance, we have o-invariance and lying down when the state is unobserved or observed. For $n = 2$, we have affinity.*

Proof: We first consider $n = 2$.

First, observe that utility does not depend directly on the drawn state ω.

Claim 1: Fixing an equilibrium, either all types report r_1, all types report r_2 or there exists one unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.

Consider the case where some individuals give either report. Then by continuity there must be a unique type, $\bar{\theta}^{GA}$, which is indifferent between the two reports. Analogous to the previous proofs this type must be unique.

Claim 2: An equilibrium exists.

This is by standard continuity arguments.

Claim 3: We observe f-invariance.

By Claim 1, if we have a unique indifferent type, then it must be 0-mass. Since all other individuals have strict preferences, and utility does not depend on the drawn state (and hence does not depend on F), the distribution of reports does not depend on F. Thus the set of equilibria will not change with F.

Although there may be multiple equilibria, we can still make predictions regarding the effect of $\hat{G}$.

Claim 4:We observe affinity.

γ has a minimum when $r = \bar{r}$. Suppose $\hat{g}(r_2)$ increases and so the induced $\bar{r}$ increases. Observe that $r_1 \leq \bar{r} \leq r_2$. Thus, when $\bar{r}$ increases, $|r_1 - \bar{r}|$ increases and $|r_2 - \bar{r}|$ decreases. So, $\gamma(r_2 - \bar{r})$ decreases, while $\gamma(r_1 - \bar{r})$ remains the same (and equal to 0). So the utility from reporting r_2 has increased, and the utility of reporting r_1 stays the same for any given individual. Therefore, more individuals will choose to report r_2. Intuitively, if players believe that there is a higher average report, then they will also believe that the audience will be less disappointed by a higher report.

Claim 5: The model exihibits o-invariance and will exhibit downwards lying regardless of observability.

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 disappointing the audience, they may lie down (in order to avoid guilt). In fact, in any equilibrium with full support on the reporting distribution, we must have some individuals lying down. Since individuals' utility only depends on their report and not their drawn state, generically individuals (other than the zero mass of individuals who are indifferent between reports) with the same parameter θ^{GA} must take the same action. Since we have full support in the reporting distribution, there is some interval of types $[\hat{\theta}^{GA}, \tilde{\theta}^{GA}]$ that strictly prefer to report r_1 over all other reports. Because F features full support, at least some individuals who have $\theta^{GA} \in [\hat{\theta}^{GA}, \tilde{\theta}^{GA}]$ must have drawn $\omega > \omega_1$.

Turning to n states, observe that the reasoning for the f-invariance result is exactly the same (because the set of indifferent types is measure 0, and utility does not depend on the drawn state).

Claim 6: Depending on parameters, we may have affinity, aversion or $\hat{g}$-invariance.

We've already presented an example of affinity for $n = 2$. We now present an example of aversion. Suppose $n = 3$, and $r_1 = \omega_1 = 0$, $r_2 = \omega_2 = 1$, $r_3 = \omega_3 = 2$.

Suppose that utility is equal to $r - \theta^{GA}\gamma(r - \bar{r})$. We now construct a cost function that is a continuous and strictly increasing approximation of the following function: $\gamma(r - \bar{r}) = 0$ for $r - \bar{r} \leq 0.6$, $\gamma(r - \bar{r}) = 3$ otherwise. Thus, we set $\gamma(0) = 0$. Then γ increases (in a continuous fashion) so that for a very small δ, when $r - \bar{r} = 0.6 - \delta$, $\gamma(r - \bar{r}) = \epsilon$ (for a very small ϵ). At that point ς increases to 3 at $r - \bar{r} = 0.6$, and then ς asymptotes to $3 + \epsilon$ as $r - \bar{r} \to \infty$. Moreover, suppose that as a limit case 10% of individuals have $\theta^{GA} = 0.5$, and the rest have $\theta^{GA} = 1$. Suppose $\hat{G^A}$ is such that $\bar{r} = 0.2$. For small enough ϵ and δ the former type of individuals reports $r_3 = 2$, the latter type reports $r_1 = 0$ (since reporting $r_1 = 0$ gives an utility of approximately 0, reporting $r_2 = 1$ gives approximately $1 - 3\theta^{GA}$, and reporting $r_3 = 2$ gives approximately $2 - 3\theta^{GA}$). Now if we shift the beliefs about the reporting distribution so that $\hat{G}^B$ implies that $\bar{r} = 0.5$, then the former type reports $r_2 = 1$ and the latter type reports $r_2 = 1$ as well (since reporting $r_1 = 0$ gives approximately 0, reporting $r_2 = 1$ gives approximately 1, and reporting $r_3 = 2$ gives approximately $2 - 3\theta^{GA}$). So we have aversion. By continuity, we can also have $\hat{g}$-invariance. □

B.7 Choice Error

One potential explanation for the observed pattern of non-maximal reports is that individuals' utility function only incorporates material payoffs, but individuals simply make mistakes when choosing, and so sometimes do not actually make the utility-maximizing report. The related Luce (1959) and McFadden et al. (1973) models of discrete choice with errors are very common specifications. This supposes that individuals have a standard utility function, but make errors when taking their action. Specifically, the utility of report r is $\phi(r)$ where ϕ is a positive function strictly increasing in r, i.e., every individual prefers to make the highest report. However, individuals do not always choose the utility maximizing report. Instead, the probability of choosing report r_i is $\frac{e^{\phi(r_i)\theta^{CE}}}{\sum_{j=1}^{n} e^{\phi(r_j)\theta^{CE}}}$. θ^{CE} is a parameter that governs the amount of "randomness" for a given individual. As θ^{CE} goes to infinity, the individual always chooses the utility maximizing report. As θ^{CE} goes to 0, reports are made with uniform chance.[45]

Proposition 9 *Suppose individuals' choices follow the Choice Error model. For arbitrary n, we have f-invariance, $\hat{g}$-invariance, o-invariance and lying down when the state is unobserved or observed.*

Proof: Observe that the chosen report does not depend on the drawn state, others' reports, or observability for any n, and we thus obtain $f-$, $\hat{g}-$ and $o-$invariance. Moreover, all individuals, conditional on a type, have the same distribution of reports regardless of the drawn state, so we observe lying down.□

B.8 Kőszegi-Rabin + LC

Kőszegi and Rabin (2006) suggest a widely used model of expectations-based reference-dependence in which the recent rational expectations serve as the reference point. We can

[45] To bring this model in line with our general theoretical framework outlined in Section 2, which is based on error-free utility maximization, one could interpret the choice error as coming from a shock to $\phi(r)$ which makes a subject prefer a particular non-maximal report. This shock would be distributed such that the choice probabilities are as in the formula in the text.

combine the intuition of the Kőszegi-Rabin model with the lying cost model. Garbarino et al. (forthcoming), in a concurrent paper, suggest and test a related 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

$$\phi(r, \omega, a; \theta^{LC}, \theta^{LAweight}, \theta^{LAmoney}, \theta^{LAcost}) = \hat{\phi}(r, c(r, \omega); \theta^{LC}) +$$
$$\theta^{LAweight}[\sum_k \theta^{LAmoney\mathbb{I}}|(u(r)-u(a(\omega_k)))|f(\omega_k)+\sum_k \theta^{LAcost\mathbb{I}}|(c(a(\omega_k), \omega_k)-c(a(\omega), \omega))|f(\omega_k)]$$

Four elements of $\vec{\theta}$ affect utility in this model. θ^{LC} parameterizes the cost of lying. $\theta^{LAweight}$ parametrizes the weight on gain-loss utility, and $\theta^{LAmoney}$ and θ^{LAcost} represent the separate gain-loss parameters for money and lying costs. $\theta^{LAmoney\mathbb{I}}$ and $\theta^{LAcost\mathbb{I}}$ are indicator functions that take on values of 1 if the argument inside the attached absolute value is positive, and $\theta^{LAmoney}$ or θ^{LAcost} respectively otherwise.

$\hat{\phi}$ takes on all the attributes that ϕ does in the LC model, and c has the exact same properties. $a(\omega_k)$ is the action that an individual expected to take, conditional on drawing ω_k. Our solution concept is the preferred personal equilibrium notion introduced in Kőszegi and Rabin (2006). A personal equilibrium a is a mapping such that if a maps $\hat{\omega}$ to $\hat{r}$, then the argmax of $\phi(r, \hat{\omega}, a; \theta^{LC}, \theta^{LAweight}, \theta^{LAmoney}, \theta^{LAcost})$ is $\hat{r}$. A personal equilibrium will exist for the reasons outlined in Kőszegi and Rabin (2006) and Kőszegi and Rabin (2007). As pointed out by those papers, 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 $\theta^{LAmoney}$ and θ^{LAcost}. We will suppose, in line with Kőszegi and Rabin (2006) and Kőszegi and Rabin (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.

Proposition 10 *Suppose individuals have Kőszegi-Rabin + LC utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we have $\hat{g}$-invariance, o-invariance and no lying down when the state is unobserved or observed.*

Proof: We first consider $n = 2$.

Claim 1: No individual lies down in any personal equilibria.

Doing so would incur lying costs and reduce monetary payoffs as well as weakly increase loss utility (decrease gain utility).

Claim 2: Conditional on a personal equilibrium, either all types report r_1, all types report r_2 or there exists a unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.

The existence and uniqueness follow from the same reasoning as in the LC model.

Claim 3: A preferred personal equilibrium exists.

Kőszegi and Rabin (2006) footnote 13 (p. 1145) shows this must be true.

Claim 4: Depending on parameters, we may have drawing in, drawing out or f-invariance.

For example, suppose as a limit case that an individual exhibits only gain-loss utility in the monetary dimension, but not in the lying cost dimension. Then an increase in $f(\omega_2)$ will increase expectations of monetary payoff, and so, conditional on drawing ω_1, an individual will be more likely to report r_2. 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. By continuity, we can generate f-invariance.

Claim 4: The model exibits $\hat{g}$-invariance.

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 $\hat{G}$. We thus obtain $\hat{g}$-invariance.

Claim 5: The model exihibits o-invariance and no downwards lying regardless of observability.

As in the LC model observability will not affect reports.

The ambiguous results on shifts in f clearly must hold for n states if it holds for two. The result on $\hat{g}$-invariance also does not depend on the number of states. □

C Models that do not Match the Findings of the Meta Study

C.1 Standard Model and Lexicographic Lying Costs

The typical assumption in economics is that in anonymous, one-shot interactions, individuals will simply maximize material payoffs, so utility is only a function of r:

$$\phi(r)$$

where utility is (strictly) increasing in r. This model cannot explain the findings of the meta study.[46]

Proposition 11 *Suppose individuals have standard utility. Then all individuals give the highest report.*

Proof: Since individuals maximizing utility implies maximizing the report, all individuals always give the highest report.□

This proposition contradicts Finding 2 of the meta study. Several papers (e.g., Demichelis and Weibull 2008, Ellingsen and Östling 2010, Navin Kartik et al. 2014) assume that individuals have weak (or lexicographic) preferences for truth-telling, i.e., individuals care about r and receive an additional small utility $\varepsilon > 0$ when they report truthfully. Since reports in our setup always yield different monetary payoffs, this model makes the same predictions as the standard model.

C.2 Reputation for Honesty

Many authors have found it plausible that individuals care about some kind of reputation that is linked to the belief of the audience player about whether the individual reported truthfully, where the audience can only observe the report but not the true state. Individuals suffer a disutility from the stigma of being perceived as a liar. One might imagine that this

[46] Moreover, the standard model predicts f-invariance, $\hat{g}$-invariance, o-invariance, and no lying down when the state is unobserved or observed.

"stigmatization aversion" is the sole reason motivating an aversion to lying. Thus, this type of model is like the Reputation for Honesty + LC model described in the body of the paper, but where θ^{LC} is always 0. Therefore, an aversion to lying is motivated solely by concerns about the beliefs of the audience. As before, because the audience's beliefs enter the utility of subjects, understanding such a model requires using the framework of Battigalli and M. Dufwenberg (2009). M. Dufwenberg and M. A. Dufwenberg (2018) introduce a similar model, but where others' beliefs about the degree of over-reporting matter for utility.

We find that such a model cannot explain the findings of the meta study. Formally, we suppose that in a Reputation for Honesty model individuals' utility is

$$\phi(r, \Lambda(r); \theta^{RH})$$

$\Lambda(r)$ is the fraction of liars and, as in the Reputation for Honesty + LC model, is the audience player's belief about whether an individual reporting r is a liar. The only element of $\vec{\theta}$ that affects utility is the scalar θ^{RH} which governs the weight that an individual applies to the stigma of being perceived as a liar. We assume ϕ is strictly increasing in its first argument and decreasing in the second argument; strictly when $\theta^{RH} > 0$. 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 (weakly) decreasing in θ^{RH} fixing the first two arguments, and that the cross partial of ϕ with respect to $\Lambda(r)$ and θ^{RH} is strictly negative, while other cross partials are 0. An equilibrium will exist because of standard continuity arguments, but because of the dependence of utility on other's strategies (via the audience's beliefs) it may not be unique.

One can show that with two states the fraction of liars at the high report is $\Lambda(r_2) = \frac{H(\bar{\theta}^{RH})f(\omega_1)}{H(\bar{\theta}^{RH})f(\omega_1)+H(\bar{\theta}^{RH})[1-f(\omega_1)]} = f(\omega_1)$. Similarly, we can show that $\Lambda(r_1) = f(\omega_2)$. This implies directly that if $f(\omega_1) \leq f(\omega_2)$ then in an equilibrium with full support the fraction of liars at r_2 would be weakly smaller than the fraction of liars at r_1. And so by saying r_2, individuals would receive both a higher monetary payoff and a weakly lower reputational cost. Thus, all individuals should say r_2 and there cannot be an equilibrium with full support, contradicting Finding 2 (when restricted to binary states) of the meta study. This result generalizes to a setting with many states as we show in the proof.[47]

Proposition 12 *Suppose individuals have Reputation for Honesty utility and F is uniform. Then all individuals give the same report.*

Proof: We first show the result for binary states and then generalize to an arbitrary number of states. Observe that utility does not depend directly on the drawn state ω.

Claim 1: Fixing an equilibrium, either all types report r_1, all types report r_2 or there exists one unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.

The reasoning is analogous to that provided for the Inequality Aversion model.

The optimal report of an individual does not depend on ω (other than for the 0-mass of indifferent individuals)

Claim 2: An equilibrium exists.

An equilibrium will exist given the continuity of ϕ and the fact that Λ is continuous in the cutoff $\bar{\theta}^{RH}$ (although it may be a corner equilibrium without full support on all reports).

By Claim 1, conditional on drawing a particular state, individuals will follow a threshold rule — people with $\theta^{RH} \geq \bar{\theta}^{RH}$ will give one report, and everyone else a different report. Suppose we have an equilibrium where a positive measure of individuals with state ω_1 report r_1. This means that there exists a set of θ^{RH}s with positive measure that strictly prefer reporting r_1 conditional on drawing ω_1. Thus the exact same set of θ^{RH}s strictly prefer reporting r_1 conditional on drawing ω_2 (since the set of indifferent types must have 0 mass).

Since the threshold is independent of the drawn state for all but a 0-mass of individuals it follows that

$$\Lambda(r_2) = \frac{H(\bar{\theta}^{RH})f(\omega_1)}{H(\bar{\theta}^{RH})f(\omega_1) + H(\bar{\theta}^{RH})[1 - f(\omega_1)]} = f(\omega_1)$$

Thus the probability of a report of r_2 being made by a liar is equal to the probability of having drawn ω_1. Similarly,

$$\Lambda(r_1) = \frac{(1 - H(\bar{\theta}^{RH}))f(\omega_2)}{(1 - H(\bar{\theta}^{RH}))f(\omega_2) + (1 - H(\bar{\theta}^{RH}))[1 - f(\omega_2)]} = f(\omega_2) = 1 - f(\omega_1)$$

[47] Moreover, the Reputation for Honesty model predicts (for $n = 2$) drawing in, $\hat{g}$-invariance, o-shift, lying down when the state is unobserved, and no lying down when the state is observed.

Thus the probability of a report of r_1 being made by a liar is equal to the probability of having drawn ω_2.

Claim 3: The equilibrium is unique.

Because there must be only a single indifferent type the equilibrium is unique.

Claim 4: With a uniform distribution we cannot have an equilibrium with full support.

If $f(\omega_1) \leq 1 - f(\omega_1)$ then the equilibrium will not have full support, i.e., not all reports occur with positive probability, since $\phi(r_1, 1 - f(\omega_1); \bar{\theta}^{RH}) < \phi(r_2, 1 - f(\omega_1); \bar{\theta}^{RH}) < \phi(r_2, f(\omega_1); \bar{\theta}^{RH})$ for any possible threshold. In other words, the utility from giving the low report must be lower than the utility of reporting the high report for any threshold.

We now turn to considering n states.

First, observe that fixing an equilibrium for any pair of states n, m there will be a unique threshold value $\bar{\theta}^{RH}_{n,m}$ for the same reasoning as in Claim 1. Similarly, by continuity an equilibrium must exist.

Consider two states, $\omega < \omega'$ along with corresponding reports $r < r'$ and suppose an equilibrium exists where $g(r) > 0$ and $g(r') > 0$. In this, denote Θ_r as the set of types willing to report r. Observe that the proportion of liars at r is then

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

By analogous reasoning, the proportion of liars at $r' = \omega'$ is $1 - f(\omega' = r')$.

Claim 5: With a uniform distribution we cannot have an equilibrium with full support.

Whenever there is an $\omega < \omega'$ such that $f(\omega) \leq f(\omega')$ 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. Thus, with a uniform distribution, all individuals will make the same report. Because the off-equilibrium beliefs are not restricted, this may not be the highest report (i.e., everyone may report r_1 in equilibrium). This may be, e.g., because the off-equilibrium beliefs imply that the subject must be a liar if they make any other report, an increase in the monetary payoff is not enough to compensate for the decreased reputation.□

C.3 Audit Model

The Audit model builds on the intuition of the Reputation for Honesty model but with a twist. Individuals' utility depends on the beliefs of the audience about whether they are a liar or not, but only in the circumstance that they actually lied up. The model captures the intuition of audits: individuals fear to be "found out" as liars. The probability of being found out depends on the 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: individuals care about the belief of the audience player about whether they reported truthfully – 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. If the audit is an actual concern about the researcher, then one can alleviate such concerns, e.g., by conducting the experiment over the phone. Our meta study, however, finds no difference in behavior when the experiment is conducted remotely (see Table A.2). 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 (2017) also discuss such an intuition for lying aversion. Because utility (potentially) depends on the audience player's belief we again use the framework of Battigalli and M. Dufwenberg (2009). Moreover, because the audience's beliefs in equilibrium must be correct, we can represent them as $\Lambda(r)$.

Using the audit intuition, individuals are "investigated" with a probability that is increasing in the audience's belief that they lied, which in equilibrium, is equal to $\Lambda(r)$, i.e., the proportion of liars that report the same r as the individual. If an individual is investigated, and discovered to have been lying upwards they face a utility cost (we suppose here that it is a fixed cost, but with binary states it is equivalent to supposing the cost depends on 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

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

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. $\Lambda(r)$ is the fraction of liars at r, which is in turn the posterior belief of the audience about the probability the individual has lied. The only element of $\vec{\theta}$ that affects utility is θ^{Aud} which governs the weight that an individual applies to the reputational cost. We assume that ϕ is strictly increasing in the first argument, decreasing in the second argument, strictly so if $\theta^{Aud} > 0$, and (weakly) decreasing in θ^{Aud}. Similarly to previous models, the cross partial of the second argument and θ^{Aud} is strictly negative, while other cross partials are 0. An equilibrium will exist because of standard continuity arguments, but because of the dependence of utility on others' strategies (via the audience's beliefs) it may not be unique.

The model fails to capture the findings of the meta study because under some circumstances it predicts that only one report is made with positive probability in equilibrium, contradicting Finding 2.[48]

Proposition 13 *Suppose individuals have Audit utility. Then there exists a distribution in $\mathcal{F}$ that induces a G in which only one state is reported.*

Proof: Fix the value of the parameters of the Audit model and suppose there are only two states/reports. For any value of $\theta^{Aud} \leq \kappa^{Aud}$ there exists some finite fraction of liars at r_2, $\Lambda^{*(\theta^{Aud})}(r_2)$, such that the value of being thought of as telling the truth and receiving r_1 is equal to the value of receiving r_2 and being thought of as a liar with probability $\Lambda^{*(\theta^{Aud})}(r_2)$: $\phi(r_1, 0; \theta^{Aud}) = \phi(r_2, \Lambda^{*(\theta^{Aud})}(r_2); \theta^{Aud})$. κ^{Aud} is finite and so consider the fraction of liars at r_2 that would make the highest type indifferent between both reports: $\Lambda_2^{*(\kappa^{Aud})}$. Now, let $f(\omega_1)$ go to zero. There exists some f^* such that for all $f(\omega_1) < f^*$, even if everyone who draws the low state says the high state, $\Lambda(r_2) < \Lambda^{*(\kappa^{Aud})}(r_2)$. This implies that all individuals will find it optimal to report the higher state. $\square$

[48] Moreover, the Audit model predicts (for $n = 2$) drawing in, aversion, o-shift, and no lying down when the state is unobserved or observed.

D Proofs for Results in Section 2 of the Main Paper

Proof of Proposition 1: *There exists a parameterization of the LC model, the Conformity in LC model, the Reputation for Honesty + LC model and of all other models listed in Appendix B (i.e., Inequality Aversion; Inequality Aversion + LC; Censored Conformity in LC; Reputation for Being Not Greedy; LC-Reputation; Guilt Aversion; Choice Error; and Kőszegi and Rabin + LC) which can explain Findings 1–4 for any number of states n and for any $F \in \mathcal{F}$.*

We first prove the proposition for the LC model.

LC Model: We will parameterize the LC model with the following function: $r - C\mathbb{I}_{r\neq\omega} - (\theta^{LC}+\epsilon)(r-\omega)^2$. r is the payoff from the report, C is a fixed cost of lying, $\mathbb{I}_{r\neq\omega}$ is an indicator function that takes on the value 0 if $r=\omega$ and 1 otherwise, ϵ is a positive constant, and θ^{LC} is the individual's aversion to lying. Thus, this functional form captures both a fixed and convex cost of lying. We prove the results in a series of steps.

We will first suppose that individuals can lie to any real value, rather than only integer values. As we will show, the results will not change when we consider the discrete (integer-valued) case.

Claim 1: Regardless of the number of states or the distribution F over them, for any given state ω there exists a cutoff type $\widetilde{\theta}^{LC}(\omega)$ so that for all $\theta^{LC} > \widetilde{\theta}^{LC}(\omega)$ individuals will not lie. Moreover, there exists an ϵ such that for any ω, $\widetilde{\theta}^{LC}(\omega) > \epsilon$

For an individual who draws a given ω, the utility of not lying is ω. If they lie, their optimal report is $r^* = \omega + \frac{1}{2(\theta^{LC}+\epsilon)}$. This gives utility of $\omega + \frac{1}{4(\theta^{LC}+\epsilon)} - C$. Notice that $\frac{\partial(\omega+\frac{1}{4(\theta^{LC}+\epsilon)}-C)}{\partial\theta^{LC}} < 0$. Moreover, as θ^{LC} goes to ∞, the maximum utility from lying goes to $\omega - C$, which is strictly less than the utility from not lying. Thus for a large enough κ^{LC}, there must exist a $\widetilde{\theta}^{LC}(\omega)$. Moreover, observe that the conditions just described do not depend on ω, immediately implying the existence of ϵ.

Claim 2: The model generates Finding 1 and Finding 2.

By Claim 1, the fraction of truth-tellers at each state ω is strictly bounded away from 0. This proves that G will have positive support on all reports (implying Finding 2). It also proves that the average payoff must be bounded away from the maximal payoff (Finding 1). Moreover, if individuals cannot choose any report, but only integers, then the optimal utility from lying must be bounded above by $\omega + \frac{1}{4(\theta^{LC}+\epsilon)} - C$. Thus, the results carry over since the result about the maximum utility obtained when θ^{LC} goes to ∞ still holds.

Moving on to proving that the model generates the other two findings, we explicitly suppose reports must take on the values $r_1, ..., r_n$. Given a distribution over θ^{LC} and a draw $\omega = \rho_m$, we can consider the induced distribution over reports $r_m, r_{m+1}, ...$(as individuals do not lie down in the LC model). Define $\overline{g}(\varrho|\rho)$ as the probability, conditional on drawing ρ, that ϱ is the optimal report when $n = \infty$. For any finite n, define the probability that an individual reports r, conditional on drawing ρ_m, as $\widetilde{g}_n(r|\rho_m)$ (notice $\widetilde{g}_\infty(r|\rho_m) = \overline{g}(r|\rho)$). The probability that any given report r is given is simply the sum of all the conditional probabilities over all states lower than r: $g(r) = \sum_{\rho=r_1}^{\rho=r} \widetilde{g_n}(r|\rho)$.

Claim 3: Suppose $n = \infty$. Consider two individuals who draw two different states; ρ and ρ'. The probability of wanting to report $\rho + k$, conditional on drawing ρ, is the same as the probability of wanting to report $\rho' + k$, conditional on drawing ρ': $\overline{g}(\rho + k|\rho) = \overline{g}(\rho' + k|\rho')$

Observe that an individual who draws ρ will prefer $\rho + k_1$ to $\rho + k_2$ if and only if $\rho + k_1 - C\mathbb{I}_{k_1 \neq 0} - (\theta^{LC} + \epsilon)(k_1)^2 \geq \rho + k_2 - C\mathbb{I}_{k_2 \neq 0} - (\theta^{LC} + \epsilon)(k_2)^2$ or $k_1 - C\mathbb{I}_{k_1 \neq 0} - (\theta^{LC} + \epsilon)(k_1)^2 \geq k_2 - C\mathbb{I}_{k_2 \neq 0} - (\theta^{LC} + \epsilon)(k_2)^2$. Moroever, an individual who draws ρ' will prefer $\rho' + k_1$ to $\rho' + k_2$ if and only if $\rho' + k_1 - C\mathbb{I}_{k_1 \neq 0} - (\theta^{LC} + \epsilon)(k_1)^2 \geq \rho' + k_2 - C\mathbb{I}_{k_2 \neq 0} - (\theta^{LC} + \epsilon)(k_2)^2$ or $k_1 - C\mathbb{I}_{k_1 \neq 0} - (\theta^{LC} + \epsilon)(k_1)^2 \geq k_2 - C\mathbb{I}_{k_2 \neq 0} - (\theta^{LC} + \epsilon)(k_2)^2$. Thus, $\overline{g}(\rho + k|\rho) = \overline{g}(\rho' + k|\rho')$.

Claim 3 is not necessarily true when n is finite. The next claim considers what happens for finite n. In doing so, we first want to highlight a useful fact. In the case where n is finite, suppose $\rho' > \rho$. If $\rho + k > r_n$ and so an individual drawing $\omega = \rho$ can't report k levels higher (since this would exceed the highest available report), then they also can't report k levels higher when drawing $\rho^{'}$ since $\rho' + k > r_n$.

Claim 4: Suppose $\rho + k > r_n$ and there are individuals who draw ρ who would want to report $\rho + k$ if $n = \infty$. In this case, these individuals (i) report r_n or (ii) tell the truth. Moreover, suppose an individual of a given type draws ρ and wants to report $\rho + k$ but cannot, and ends up telling the truth. If the same individual draws $\rho' > \rho$ and wants to report $\rho' + k$ but cannot, they will also end up telling the truth.

We prove the first part of the claim in two steps. First, we want to establish that this individual who wants to report $\rho + k > r_n$ must find that reporting r_n gives a higher utility than any other report $r > \rho$ (recall that individuals will never report below their draw ρ). To do so, we simply show that utility, conditional on reporting more than ρ, is falling the farther the report is from the optimal, but unavailable, report. Observe that the second derivative of the utility function for all $r > \rho$ is $-2(\theta^{LC} + \epsilon)$. This is strictly negative. Suppose the optimal report is r^*, and $|\hat{r} - r^*| \geq |r - r^*|$, where both $\hat{r}$ and r are larger than ρ. Then utility from report r is larger than the utility of reporting $\hat{r}$. In other words, the utility for an individual is lower the farther a given report is from the optimal report. Then suppose the highest report that is possible is $r_n < \infty$, and $r^* > r_n$. Then, if an individual lies, they will report r_n. Of course, it may be optimal also not to lie, in which case ρ must give maximal utility.

We prove the second part of the claim now. To do this, we suppose that, above r_n, reports could (if they were allowed) take on any value (not just the integers). Suppose an individual of a given type draws ρ and wants to report $r^* = \rho + k$ but cannot, and ends up telling the truth. From Claim 3, this indiviudal would want to report $r^* = \rho' + k$ if they drew ρ'. Given an optimal report $r^*(\rho)$ (it is a function of the drawn state, and we surpress the dependence on the individual's type) not equal to the drawn state, we know that the utility from reporting r is $r - C - (\theta^{LC} + \epsilon)(r - r^*(\rho) + \frac{1}{2(\theta^{LC}+\epsilon)})^2$. Denote the difference between any given report r and the optimum report as $d(r, \rho) = r - r^*(\rho)$.

From the previous paragraph we know that this individual will either report r_n or ρ when drawing ρ. ρ is reported if and only if $r_n - C - (\theta^{LC} + \epsilon)(d(r_n, \rho) + \frac{1}{2(\theta^{LC}+\epsilon)})^2 \leq \rho$. Moreover, observe that $d(r_n, \rho)$ is negative, and the derivative of the utility function with respect to d, so long as it is negative, is positive.

If the same individual draws $\rho' > \rho$ we know that this individual will either report r_n or ρ' when drawing ρ'. $d(r_n, \rho')$ is negative and it is more negative than $d(r_n, \rho)$: $d(r_n, \rho') \leq d(r_n, \rho) \leq 0$. This implies that the utility of reporting r_n, having drawn ρ', $r_n - C - (\theta^{LC} + \epsilon)(d(r_n, \rho') + \frac{1}{2(\theta^{LC}+\epsilon)})^2$ must be less than the utility of reporting r_n, having drawn ρ, $r_n - C - (\theta^{LC} + \epsilon)(d(r_n, \rho) + \frac{1}{2(\theta^{LC}+\epsilon)})^2$. Moreover, the utility of reporting ρ', having drawn ρ', is larger than the utility of reporting ρ, having drawn ρ. Thus $\rho' \geq r_n - C - (\theta^{LC} + \epsilon)(d(r_n, \rho') + \frac{1}{2(\theta^{LC}+\epsilon)})^2$, and so this individual will want to report the truth.

Claim 5: The probability, conditional on drawing ρ, of telling the truth (i.e. reporting the drawn state), is increasing in ρ.

To see this, consider the same individual who could have either drawn ρ or $\rho' > \rho$. There are two cases. First, suppose that for this individual the optimum, when $n = \infty$, after drawing

ρ is to say $\rho + k$. Moreover, $\rho + k < r_n$. In this case, the individual actually reports $\rho + k$. We showed above (Claim 3) that the individual would then like to report $\rho' + k$ when drawing ρ'. If they are able to do so, then they will. But it is possible that $\rho' + k > r_n$. Therefore, the unconstrained optimal report is not available. As shown in Claim 4, such individuals may report r_n, but may also report ρ'. Thus aggregating across individuals, in this case we observe a higher chance of reporting ρ', conditional on drawing ρ' than reporting ρ, conditional on drawing ρ.

In the second case, suppose the optimum $\rho + k$ is greater than r_n. Then, as we have shown in the paragraph previous to the statement of Claim 5, there is a higher chance of telling the truth conditional on drawing $\rho' > \rho$ (relative to drawing ρ).

The preceding two paragraphs imply that the outflow of individuals (i.e. individuals who drew a state but do not give the corresponding report) is decreasing in the state ρ, conditional on having drawn that state. Thus, if there is the same chance of drawing any given state, the outflows must be decreasing in ρ.

Claim 6: The probability, conditional on drawing a state lower than r, that r is the optimal report, is increasing in r.

Another way of stating Claim 6 is that conditional on drawing a state $\omega \leq r$, the fraction of individuals who find r the optimal report is increasing in r. To see this, first consider some $r < r_n$. For any individual giving a report r who is lying, it has to be the case that they drew ρ and $r = \rho + k$ was the optimal report to give. We have previously shown (Claim 3) that this implies that this same individual would report $r - 1$ if they drew $\rho - 1$. If $\rho - 1 \geq \omega_1$ then this happens. But if $\rho - 1 < \omega_1$ then there are no individuals who could have drawn $\rho - 1$, and so the set of people lying to $r - 1$ must be smaller than the set of people lying to r, when $r < r_n$. Observe that this reasoning is also true for individuals who lie to r_n, conditional on those individuals having r_n as the optimal report even if it were possible to report $r_n + 1$. However, there are also individuals who are lying to r_n because they cannot report any higher than r_n. Thus, the number of people lying at r_n is larger than at r_{n-1}. This implies that so long as there was the same chance of drawing all states, the inflows of individuals (i.e. individuals who give a report but did not draw the corresponding state) is increasing in the state ρ.

Claim 7: The model generates Finding 3.

Since for uniform distributions outflows are decreasing in the state (and corresponding report) but inflows are increasing, $g(r)$ must be increasing (Finding 3).

Last we need to show that some state, other than the highest, is over-reported for all allowable distributions with more than 3 states (Finding 4).

Claim 8: Over-reporting occurs for the second highest state when F is uniform.

First, calibrate the model so that no individuals are willing to report more than two states/reports higher than what they drew. This means we find values of C and ϵ so that the individuals with the lowest costs of lying are willing to lie up 2, but not 3 reports. In other words, C and ϵ have values so that $\rho + 2 - C - 4\epsilon > \rho$ and $\rho + 3 - C - 9\epsilon < \rho$ or $2 > C + 4\epsilon$ and $3 < C + 9\epsilon$. Individuals who drew ω_j will thus report either ω_j, ω_{j+1} or ω_{j+2}. Moreover, individuals who desire to report ω_{j+2}, but cannot (i.e. those individuals who drew ω_n or ω_{n-1}), simply do not lie (because of the fixed cost). With more than 3 states and a uniform distribution, the second highest state must be over-reported. To see this, observe that the only people who may report the highest and second highest states are individuals who drew one of the top four states. Moreover, $\overline{g}(r_n|\omega_{n-1}) = \widetilde{g}(r_n|\omega_{n-1})$ since those that drew ω_{n-1} and would like to report r_{n+1}, but obviously cannot, end up reporting r_{n-1}. This reasoning extends, so that $\overline{g}(r_n|\omega_{n-1}) = \overline{g}(r_{n-1}|\omega_{n-2}) = \overline{g}(r_{n-2}|\omega_{n-3}) = \widetilde{g}(r_{n-1}|\omega_{n-2}) = \widetilde{g}(r_{n-2}|\omega_{n-3})$. Moreover $\overline{g}(r_n|\omega_{n-2}) = \overline{g}(r_{n-1}|\omega_{n-3}) = \widetilde{g}(r_n|\omega_{n-2}) = \widetilde{g}(r_{n-1}|\omega_{n-3})$. Thus the inflows to r_{n-1} are $\frac{1}{n}\overline{g}(r_{n-1}|\omega_{n-2}) + \frac{1}{n}\overline{g}(r_{n-1}|\omega_{n-3})$. By construction the outflows from ω_{n-1} are $\frac{1}{n}\overline{g}(r_n|\omega_{n-1})$. This implies that the outflows are smaller than the inflows, so the state must be overreported.

Claim 9: Over-reporting occurs for the second highest state for any distribution in $\mathcal{F}$.

Finally, consider any distribution in $\mathcal{F}$ with 3 or more states. Then the inflows to ω_{n-1} are $f(\omega_{n-2})\overline{g}(r_{n-1}|\omega_{n-2}) + f(\omega_{n-3})\overline{g}(r_{n-1}|\omega_{n-3})$ and the outflows are $f(\omega_{n-1})\overline{g}(r_n|\omega_{n-1})$. Since $f(\omega_{n-1}) \leq f(\omega_{n-2})$ the inflows must exceed the outflows.

The series of claims thus proves the LC model can match Findings 1–4 of the meta study.

We next turn to the models that limit to the LC model: The Reputation for Honesty + LC model, the LC-Reputation model, the Conformity in LC model, the Inequality Aversion + LC model, the Censored Conformity in LC model and the Kőszegi-Rabin+LC model (for details of these models, see Section 2 and Appendix B). Because of our construction of these models, they do not formally nest the LC model. Instead, they limit to the LC model in various ways. The Reputation for Honesty + LC model, the LC-Reputation model, the Inequality Aversion + LC model and the Kőszegi-Rabin + LC model limit to the LC model as the distribution on the $\theta \neq \theta^{LC}$ converges to 0. For these models, it is clear that as the other cost components become negligible, behavior will be almost entirely governed by the LC cost component. The Conformity in LC and Censored Conformity in LC models limit to the LC model as η becomes a function that does not depend on its second argument. Again, this implies that individuals' cost of lying no longer depends on others' actions, giving us behavior arbitrarily close to the LC model. Thus, they can also explain Findings 1–4.

We now turn to the remaining models.

The Inequality Aversion Model (see Appendix B.1): Suppose as a limiting case, we have 60% of individuals who simply maximize monetary payoff and 40% who experience an infinite loss of utility if they are above the mean report, but no loss if they are below. Then for any number of reports/states there exists an equilibrium where 60% of individuals report r_n, and 40% report r_{n-1}.

We show that this is an equilibrium in two steps. First, in any equilibrium the former type of individuals always give the highest report. Second, in the equilibrium we are constructing, observe that the mean report lies between r_{n-1} and r_n. Thus, the second type of player experiences an infinite loss of utility if they give report r_n, but experience utility r if they given any report $r < r_n$, and so they report r_{n-1}.

We show that this equilibrium has the desired properties. More than one report is given with positive probability, the average payoff is bounded away from the maximum payoff, and the histogram is (weakly) increasing. With any uniform F with more than 3 states, a non-maximal report (the second highest report) is made more often than its true likelihood. The equilibrium reporting distribution doesn't depend on F, and any other allowable F places lower weight on the second highest state than a uniform distribution and so we also have over-reporting for all $F \in \mathcal{F}$ with more than 3 states. Of course there are also other potential equilibria, but we just focus on the one with desired properties. Thus, this distribution of reports matches Findings 1–4.

The Reputation for Being Not Greedy Model (see Appendix B.4): To prove that the model can match the findings, assume that $\phi(r) = \theta^{RNG} E[\theta^{RNG}|r] + (1 - \theta^{RNG})r$ and a distribution of θ^{RNG} where in the limit there are two types. The first type has $\theta^{RNG} = 0$, thus cares nothing at all for reputation and only about material payoffs. They always report r_n. The second type has $\theta^{RNG} = -\frac{1}{2} + \frac{\sqrt{5}}{2}$. We propose an equilibrium where this type reports

r_{n-1}. For any individual of the second type in this equilibrium the utility from reporting the highest report is $0 + (1 - \theta^{RNG}_{High})r_n = (1 - \theta^{RNG}_{High})r_n$, the utility of the second highest report is $\theta^{RNG}_{High}\theta^{RNG}_{High} + (1 - \theta^{RNG}_{High})r_{n-1}$. Setting these equal and solving the quadratic equation $0 = (\theta^{RNG}_{High})^2 + \theta^{RNG}_{High} - 1$ gives $\theta^{RNG} = -\frac{1}{2} + \frac{\sqrt{5}}{2}$. Thus the high types are indifferent between reporting r_n and r_{n-1} and we assume they report r_{n-1}. Thus, this is an equilibrium. Suppose the type that doesn't care at all about reputation composes 60% of the population, and the rest is the higher type. As described for the Inequality Aversion model above, this distribution of reports matches Findings 1–4.

The Guilt Aversion Model (see Appendix B.6): To see that a model of guilt aversion can match the meta-study findings, we do a construction analogous to the Inequality Aversion model. Suppose as a limiting case that 60% of individuals simply maximize monetary payoff. The remaining 40% of individuals experience an infinite loss of utility if they disappoint the audience player. Then for any number of reports/states there exists an equilibrium where 60% of individuals report r_n and 40% report r_{n-1}. We show that this is an equilibrium in two steps. First, in any equilibrium the former type of individuals always give the highest report. Second, in the equilibrium we are constructing, observe that the audience expects a report between r_{n-1} and r_n. Thus, the second type of player experiences an infinite loss of utility if they give report r_n, but experiences utility r if they given any report $r < r_n$, and so they report r_{n-1}. As described above, this distribution of reports matches Findings 1–4.

The Choice Error Model (see Appendix B.7): Since ϕ is always finite, so long at $\theta^{CE} < \infty$ the Choice Error model predicts that more than one report is made with positive probability and that the payoffs are bounded away from the maximum payoff. Moreover g is strictly increasing by construction. The last thing to prove is that we get over-reporting of a non-maximal report when $n > 3$. We will construct a ϕ so that the second highest state is always reported with probability more than $\frac{1}{n}$ which will satisfy this condition. To simplify matters, assume a limit case: that all individuals have the same type $\theta^{CE} = 1$. We denote $\hat{\phi}(r) = e^{\theta^{CE}\phi(r)}$. Let $\hat{\phi}(r_1) \to 0$ and allow for $\hat{\phi}(r_2)$ to be any particular value. We construct our result inductively showing that we can generate over-reporting of a non-maximal report for any $n \geq 3$. If we have three outcomes, then we need: $\frac{\hat{\phi}(r_2)}{\hat{\phi}(r_1)+\hat{\phi}(r_2)+\hat{\phi}(r_3)} > \frac{1}{3} \iff 3\hat{\phi}(r_2) > \hat{\phi}(r_1)+\hat{\phi}(r_2)+\hat{\phi}(r_3) \iff 2\hat{\phi}(r_2) > \hat{\phi}(r_3)$. We can choose any value of $\hat{\phi}(r_3)$ that satisfies this bound (and is greater than $\hat{\phi}(r_2)$). If we consider instead four reports, then it must be that

$\frac{\hat{\phi}(r_3)}{\hat{\phi}(r_1)+\hat{\phi}(r_2)+\hat{\phi}(r_3)+\hat{\phi}(r_4)} > \frac{1}{4}$, or $4\hat{\phi}(r_3) > \hat{\phi}(r_1) + \hat{\phi}(r_2) + \hat{\phi}(r_3) + \hat{\phi}(r_4) = \hat{\phi}(r_2) + \hat{\phi}(r_3) + \hat{\phi}(r_4)$, or $3\hat{\phi}(r_3) - \hat{\phi}(r_2) > \hat{\phi}(r_4)$. We then choose a value of $\hat{\phi}(r_4)$ that satisfies this constraint, and is greater than $\hat{\phi}(r_3)$. One can iterate the bounds inductively so that for the nth report, we can choose a $\hat{\phi}(r_n)$ such that $(n-1)\hat{\phi}(r_{n-1}) - \sum_{j=1}^{n-2} \hat{\phi}(r_j) > \hat{\phi}(r_n) > \hat{\phi}(r_{n-1})$. Observe that the reporting distribution generated in our construction doesn't depend on F, and any other allowable F places lower weight on the second highest state than a uniform distribution and so we have over-reporting for all $F \in \mathcal{F}$ with more than 3 states. □

Proof of Proposition 2:

- *Suppose individuals have LC utility. For an arbitrary number of states n, we have f-invariance, $\hat{g}$-invariance, o-invariance and no lying down when the state is unobserved or observed.*

- *Suppose individuals have Conformity in LC utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we may have affinity, aversion or $\hat{g}$-invariance, we have o-invariance and no lying down when the state is unobserved or observed. For $n = 2$, we have drawing out when the equilibrium is unique and we have affinity.*

- *Suppose individuals have Reputation for Honesty + LC utility. For arbitrary n, depending on parameters, we may have drawing in, drawing out or f-invariance, we may have affinity, aversion or $\hat{g}$-invariance, we have o-shift, depending on parameters, we may have lying down or not when the state is unobserved, and we have no lying down when the state is observed. For $n = 2$, we have drawing in when the equilibrium is unique.*

We first prove an initial lemma.

Lemma 1 *For all models, the results regarding o-shift/o-invariance and regarding lying down do not depend on the number of states.*

Proof of Lemma 1: For models that have o-shift, the shift occurs because if the audience player has information about the state, it changes their beliefs about the subject and this affects the subject's utility. This occurs regardless of the number of states. For models that have o-invariance, the audience's knowledge of the state does not change a player's utility. This again is unrelated to the number of states.

For models that can feature lying downward, there are three cases. First, in the Inequality Aversion, Guilt Aversion, and Choice Error model, the report does not depend on the true state and since there is full support on states and reports, we always have downwards lying irrespective of the number of states.

Second, in the Reputation for Honesty + LC, LC-Reputation and Inequality Aversion + LC models, there could be downwards lying or not for $n = 2$ and thus also for $n > 2$.

Third, for the remaining model that features lying down (Reputation for Not Being Greedy), utility depends on the audience's beliefs and lying down occurs because it can help shift these beliefs. Regardless of whether the state is observed or not, there is an incentive to possibly lie down for any number of states.

For models that do not feature lying downward (i.e., LC, Conformity in LC, Censored Conformity in LC, and Kőszegi-Rabin + LC), this happens because lying down triggers a weakly higher lying cost and leads to a lower monetary payoff relative to truth-telling. This is independent of the number of states and observability. □

When proving our results regarding the comparative statics of shifts in F and $\hat{G}$, we will prove results for an equivalent, but simpler to work with, formulation of the shifts. Rather than focusing on shifts of first order stochastic dominance which maintains the same set of support, we focus on shifts where we move weight from a single lower state to a single higher state. For example, when considering changes in F from a distribution F^A to another distribution F^B, we suppose that $f^A(\omega_i) = f^B(\omega_i)$ for all $i = 1, 2, ..., j-1, j+1, ..., k-1, k+1, ..., n$, $f^B(\omega_k) = f^A(\omega_k) + \epsilon$, and $f^B(\omega_j) = f^A(\omega_j) - \epsilon$ for some $0 < \epsilon < f^A(\omega_j)$. Any shift of this kind induces first order stochastic dominance. Moreover, by the definition of first order stochastic dominance we can decompose any shift in first order stochastic dominance on a finite distribution into a finite number of these shifts. This works analogously for shifts in $\hat{G}$. Thus we get the following (equivalent) reformulations of our definitions:

Definition 1' *Consider two pairs of distributions: F^A, G^A and F^B, G^B where G^j is the reporting distribution associated with F^j, and they all have full support. Suppose further that $f^A(\omega_i) = f^B(\omega_i)$ for all $i = 1, 2, ..., j-1, j+1, ..., k-1, k+1, ..., n$, $f^B(\omega_k) = f^A(\omega_k) + \epsilon$, and $f^B(\omega_j) = f^A(\omega_j) - \epsilon$ for some $0 < \epsilon < f^A(\omega_j)$. A model exhibits drawing in/drawing out/f-invariance if $1 - \frac{g^B(r_1)}{f^B(\omega_1)}$ is larger than/smaller than/the same as $1 - \frac{g^A(r_1)}{f^A(\omega_1)}$.*

Definition 2' *Fix a distribution over states F and consider two pairs of distributions $\hat{G}^A, G^A$ and $\hat{G}^B, G^B$, where G^j is the reporting distribution induced by F and by the belief that others will report according to $\hat{G}^j$. Moreover, suppose that all exhibit full support and that $\hat{g}^A(r_i) = \hat{g}^B(r_i)$ for all $i = 1, 2, ..., j-1, j+1, ..., k-1, k+1, ..., n$, $\hat{g}^B(r_k) = \hat{g}^A(r_k) + \epsilon$, and $\hat{g}^B(r_j) = \hat{g}^A(r_j) - \epsilon$ for some $0 < \epsilon < \hat{g}^A(r_j)$. A model exhibits* affinity/aversion/$\hat{g}$-invariance *if $g^B(r_n)$ is larger than/smaller than/the same as $g^A(r_n)$.*

To prove the rest of the results we first prove the results for binary states/reports. We do this because it allows for development of the intuitions underlying the proofs. We then prove the results for an arbitrary number of states/reports. We consider each model in turn.

LC model: First we consider $n = 2$.

Claim 1: No individual lies down.

In doing so they would pay a weakly higher lying cost and receive a lower monetary payoff than if they told the truth.

Claim 2: Conditional on drawing ω_1 either all types report r_1, all types report r_2 or there exists a unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.

We show that if neither of the first two cases holds there needs to be a unique cutoff type. Suppose that some individuals drawing ω_1 report r_1 and others report r_2. By continuity of the utility function there must be a type (cutoff type) $\bar{\theta}^{LC}$, such that $\phi(r_1, c(r_1, \omega_1); \bar{\theta}^{LC}) = \phi(r_2, c(r_2, \omega_1); \bar{\theta}^{LC})$. We can show this cutoff type will be unique. By construction $\frac{\partial^2 \phi}{\partial c \partial \theta} < 0$ and $\frac{\partial^2 \phi}{\partial r \partial \theta} = 0$. Therefore, since $\phi(r_2, c(r_2, \omega_1); \bar{\theta}^{LC}) - \phi(r_1, c(r_1, \omega_1); \bar{\theta}^{LC}) = 0$, then for all $\theta^{LC} > \bar{\theta}^{LC}$, $\phi(r_2, c(r_2, \omega_1); \bar{\theta}^{LC}) - \phi(r_1, c(r_1, \omega_1); \bar{\theta}^{LC}) < 0$ and for all $\theta^{LC} < \bar{\theta}^{LC}$, $\phi(r_2, c(r_2, \omega_1); \bar{\theta}^{LC}) - \phi(r_1, c(r_1, \omega_1); \bar{\theta}^{LC}) > 0$. Therefore, individuals with $\theta^{LC} < \bar{\theta}^{LC}$ who draw ω_1 will report r_2. Individuals with $\theta^{LC} > \bar{\theta}^{LC}$ who draw ω_1 will report r_1.

Claim 3: The model exibits f-invariance.

Given Claim 2, and the fact that no one would lie down (Claim 1), we can calculate our test statistic: $1 - \frac{g(r_1)}{f(\omega_1)} = 1 - \frac{(1-H(\bar{\theta}^{LC}))f(\omega_1)}{f(\omega_1)} = 1 - (1 - H(\bar{\theta}^{LC})) = H(\bar{\theta}^{LC})$. This condition does not depend on F.

Claim 4: The model exibits $\hat{g}$-invariance.

The fact that an individual's utility does not depend on G in any way allows us to immediately observe that it exhibits $\hat{g}$-invariance.

Claim 5: The model exihibits o-invariance and no downwards lying regardless of observability.

The lying costs in this model are internal costs and they do not depend on the inference others are making about any given person. Thus, individuals do not care whether their state was observed.

We next consider n states. We can generalize our results easily.

Observe that for each pair r_i, r_j of potential reports there is a state-conditional threshold such that an individual with that threshold would be indifferent between that pair of reports (such thresholds only exist where both reports r_i and r_j are both weakly larger than ω, since no individuals lie down): denote it $\bar{\theta}^{LC}_{r_i,r_j,\omega}$: $\phi(r_i, c(r_i, \omega); \bar{\theta}^{LC}_{r_i,r_j,\omega}) = \phi(r_j, c(r_j, \omega); \bar{\theta}^{LC}_{r_i,r_j,\omega})$. Clearly this is unique and does not depend on F as before. Denote $\bar{\theta}^{LC}_{\omega} = \min_{r_j} \bar{\theta}^{LC}_{r=\omega,r_j,\omega}$. This is the highest type that will be willing to lie, and in fact this type will be indifferent between telling the truth and lying (since it is the minimum of all the thresholds between reporting the drawn state and reporting some other state). All lower types will lie to some other state. Since no individuals lie down, then the probability of an individual giving the lowest report is $g(r_1) = H(\bar{\theta}^{LC}_{\omega_1})f(\omega_1)$. Thus, shifting the distribution above the lowest outcome doesn't change the conditional probability of someone reporting the lowest outcome. Thus we get f-invariance. Since the thresholds do not depend on G shifts in $\hat{G}$ have no effect and so we get $\hat{g}$-invariance.

Conformity in LC model: We first consider $n = 2$.

Claim 6: No individual lies down

In doing so they would pay a weakly higher lying cost and receive a lower monetary payoff than if they told the truth.

Claim 7: Fixing an equilibrium, conditional on drawing ω_1 either all types report r_1, all *types report r_2 or there exists a unique type that is indifferent between r_1 and r_2 and all types higher than that report r_1, and all others report r_2.*

In the case that some types drawing ω_1 give one report and others the other, by continuity there must be a type that conditional on drawing ω_1 is indifferent between the two reports, and so satisfies the condition $\phi(r_1, \eta(0, \bar{c}); \bar{\theta}^{CLC}) = \phi(r_2, \eta(c, \bar{c}); \bar{\theta}^{CLC})$ where c denotes the cost of lying to report r_2 (given that ω_1 was drawn). If no such type exists, then all indviduals would give the same report. As with the LC model, this type will be unique for the exact same reasoning (since fixing the equilibrium $\bar{c}$, this model is the LC model). Of course, this threshold may shift across different equilibria.

Claim 8: An equilibrium exists.

An equilibrium will exist given the continuity of ϕ and η and the fact that $\bar{c}$ is continuous in the cutoff $\bar{\theta}^{CLC}$.

However, it may not be unique. Intuitively this is true because individuals' lying behaviors are complements. To find the set of equilibria consider the function $\zeta(\bar{\theta}^{CLC})$, which maps from Θ to Θ: this will be the function whose fixed points will characterize the equilibria. Given a threshold $\bar{\theta}^{CLC}$ that all other individuals are using, $\zeta(\bar{\theta}^{CLC})$ is a function that gives the optimal threshold if there exists one in the allowed range of θ^{CLC}; it returns κ^{CLC} (the upper bound of the distribution of types) if the threshold is above the range; and gives 0 (the lower bound of the distribution of types) if the threshold is below the range. This ensures ζ maps from $[0, \kappa^{CLC}]$ to itself. It also implies, with a unique equilibrium, the graph of ζ must cross the 45-degree line from above to below. Finding the fixed point(s) of $\zeta(\bar{\theta}^{CLC})$ characterizes the equilibrium.

Claim 9: The model exhibits drawing out.

Suppose that the equilibrium is unique. Now let $f(\omega_1)$ fall. For any $\bar{\theta}^{CLC}$ as $f(\omega_1)$ falls $\bar{c}$ must fall. Thus $\zeta(\bar{\theta}^{CLC})$ must fall for all $\bar{\theta}^{CLC}$. 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_1 report r_2. Thus we observe drawing out.

Claim 10: The model exhibits affinity.

Since G enters in the utility function directly (because no one lies down and there are two states and G has thus a one-to-one mapping with $\bar{c}$) we can still make predictions regarding the effect of $\hat{G}$ even though we may not have a unique equilibrium. To see that we observe affinity, notice that fixing F, increasing $\hat{g}(r_2)$ implies that the individual believes that there are more liars. Thus the costs of lying fall, and so more individuals are willing to lie.

Claim 11: The model exihibits o-invariance and no downwards lying regardless of observability.

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.

We now turn to n states.

As mentioned for the binary world, fixing the level of lying in society, the model behaves exactly like an LC model, where among the individuals who drew ω there will be a set of thresholds that denote which state they should report. Since all types have zero measure, this implies that conditional on a value of $\bar{c}$, generically individuals have a unique best action (conditional on any drawn state). Thus, we can think of the equilibrium as simply finding a fixed point in the aggregate level of lying: $\zeta(\bar{c})$, which maps from the aggregate level of lying to itself. Because of continuity an equilibrium will always exist.

Claim 12: Depending on parameters, we may observe drawing in, drawing out or f-invariance.

We construct an example to demonstrate drawing in (since we have already shown drawing out for $n = 2$). Suppose $n = 4$. Since no one lies down, no one drawing the highest state lies. Moreover, suppose that the cost structure has two properties: (i) individuals, if they lie, lie up at most one report, and (ii) the cost of lying up one state is increasing in the drawn state. Key to the example is that there is a negligable mass of individuals who draw ω_2 who are near the threshold type (below which they report r_3, above which they report r_2). Instead, almost all individuals who draw ω_2 and lie have a strong preference for lying (i.e. the utility they obtain from reporting r_3 is much larger than the utility they obtain from reporting r_2). To obtain F^B from F^A, fix $f^A(\omega_1)$ and $f^A(\omega_4)$ and shift weight from ω_2 to ω_3. Shifting individuals to ω_3 increases their costs of lying (and reduces the benefits), but if their preference for lying up was strong enough at ω_2, then almost all of the individuals who now draw ω_3 (instead of ω_2) will continue to want to lie. Thus, $\bar{c}$ will increase. But this means that conditional on drawing

ω_1, individuals are more likely to lie, exhibiting drawing in, opposite to the prediction of the two state/report case. By continuity, it is also possible to generate f-invariance.

Claim 13: Depending on parameters, we may observe affinity, aversion or $\hat{g}$-invariance.

We have shown affinity for $n = 2$. We now demonstrate an example for aversion. Suppose that the shift in $\hat{G}$ induces a belief that $\bar{c}$ has increased (as it does in the binary case). We show that even if $\bar{c}$ has risen we may observe aversion. Let $n = 3$. First, suppose as a limit case all individuals are of the same type and utility is equal to $u(r) - \eta(c, \bar{c})$. Suppose $u(r_1) = 0, u(r_2) = 2$ and $u(r_3) = 4$, and that the cost function is such that individuals drawing ω_2 and ω_3 never want to lie. But $c(r_2, \omega_1) = 0.2$ and $c(r_3, \omega_1) = 0.4$. First, consider an equilibrium where $\eta(0.2, \bar{c}) = 1$ and $\eta(0.4, \bar{c}) = 2.8$. All individuals drawing ω_1 report r_3. Now suppose the average cost of lying rises to $\bar{c}'$ and at the new value $\eta(0.2, \bar{c}') = 0.2$ and $\eta(0.4, \bar{c}') = 2.4$. Now all individuals drawing ω_1 report r_2. Conversely, this also means that if $\bar{c}$ falls we can either observe more reporting or less reporting of the highest report. Thus, regardless of the shift in beliefs about $\bar{c}$ we may observe either affinity or aversion. By continuity, it is also possible to generate $\hat{g}$-invariance.

Reputation for Honesty + LC: We first consider $n = 2$.

Claim 14: In any equilibrium, r_2 has to have more liars.

Suppose no one lies down. Then clearly r_2 has more liars. Now suppose people do lie down, and r_2 has fewer liars than r_1. In this case, consider the individuals whose state is ω_2. They would obtain a better reputation, lower lying costs and a higher monetary payoff, by simply reporting r_2. So, no one would lie down – a contradiction. Thus, r_2 must have more liars.

Claim 15: Fixing θ^{LC} and an equilibrium, $\phi(r_2, c(r_2, \omega_1), \Lambda(r_2); \theta^{LC}, \theta^{RH}) - \phi(r_1, c(r_1, \omega_1), \Lambda(r_1); \theta^{LC}, \theta^{RH})$ is falling in θ^{RH}.

This is immediately implied by the fact that $\Lambda(r_2) > \Lambda(r_1)$ (as shown in Claim 14), $\frac{\partial \phi}{\partial \Lambda} < 0$, $\frac{\partial^2 \phi}{\partial r \partial \theta^{RH}} = 0$ and $\frac{\partial^2 \phi}{\partial \Lambda \partial \theta^{RH}} < 0$ (by our assumption of additive separability).

Similarly, fixing θ^{RH} and an equilibrium, $\phi(r_2, c(r_2, \omega_1), \Lambda(r_2); \theta^{LC}, \theta^{RH}) - \phi(r_1, c(r_1, \omega_1), \Lambda(r_1); \theta^{LC}, \theta^{RH})$ is decreasing in θ^{LC}. We can make the analogous statements about what happens conditioning instead on ω_2 being drawn.

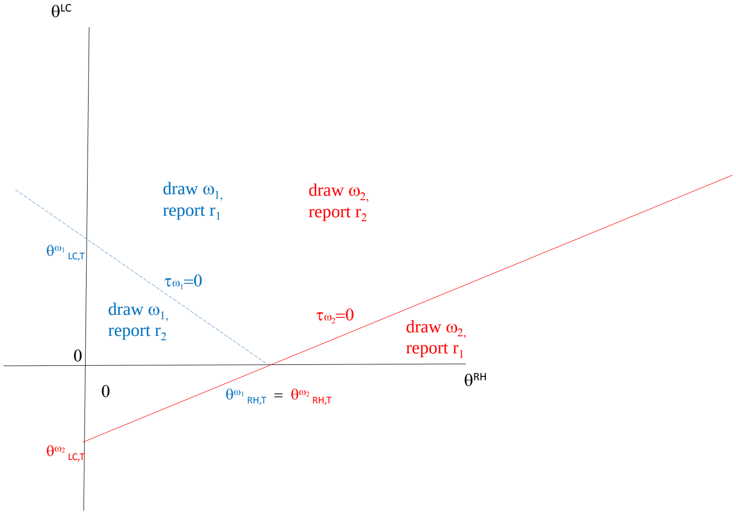
We can define a "threshold function" for each state $\tau_{\omega_i}(\theta^{LC}, \theta^{RH})$, which, given the equilibrium and an individual's given type, gives the utility of reporting $r_{i \neq j}$ versus r_i, conditional on having drawn ω_i. These are continuous functions. If τ is less than or equal to 0, the individual will report their state, otherwise they will lie.

Claim 16: Fixing θ^{LC} and an equilibrium, $\tau_{\omega_i}(\theta^{LC}, \theta^{RH})$ is equal to 0 for at most one value of θ^{RH}. Similarly fixing θ^{RH}, $\tau_{\omega_i}(\theta^{LC}, \theta^{RH})$ is equal to 0 for at most one value of θ^{LC}.

This is immediately implied by the preceding claims.

Thus, we can think of the set of indifferent individuals, i.e. the set of points where $\tau_{\omega_i}(\theta^{LC}, \theta^{RH}) = 0$, as a function in the space $\theta^{LC} \times \theta^{RH}$; or graphically, given that utility is linear in both θ^{RH} and θ^{LC}, a line in two-dimensional Euclidean space (see Figure D.1).

Figure D.1: Thresholds for Reputation for Honesty + LC Model



We know that fixing θ^{RH}, as θ^{LC} increases, individuals' relative value of reporting what they drew increases.

Claim 17: If an individual draws ω_1 and reports r_2 then an individual with the same preference parameters, but with a draw ω_2, must also report r_2. Moreover, if an individual draws ω_2 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_2 gives the same reputational value and the same monetary payoff to both individuals but the individual who drew ω_1 pays an LC cost (analogous reasoning works for the second statment).

We can characterize the equilibrium in terms of the intercepts of the threshold function, rather than the probability of being a liar. Observe that given H and a utility function, the probability that, conditional on drawing a particular state, an individual lies is characterized by $\tau_{\omega_i}(\theta^{LC}, \theta^{RH}) = 0$. Since the threshold functions $\tau_{\omega_i}(\theta^{LC}, \theta^{RH}) = 0$ are always linear in θ^{LC} and θ^{RH} they can be characterized by their θ^{LC} intercept and their θ^{RH} intercept, denoted $\theta^{\omega_i}_{LC,T}$ and $\theta^{\omega_i}_{RH,T}$. Moreover, since the LC portion of costs never depends on the distribution of responses, the $\theta^{\omega_i}_{LC,T}$ intercept (i.e. the threshold value of $\theta^{\omega_i}_{LC,T}$ when $\theta^{RH} = 0$) must always be the same. Therefore, we can think of each of the threshold "lines" (one for each drawn state) as being characterized by a single intercept: $\theta^{\omega_i}_{RH,T}$. The thresholds $\theta^{\omega_i}_{RH,T}$ (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 $\zeta(\theta^{\omega_1}_{RH,T}, \theta^{\omega_2}_{RH,T})$, which maps from the thresholds that everyone is using into best-response thresholds. This function's fixed points will characterize equilibria. Because we are looking at the θ^{RH} intercepts, the LC 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 $\theta^{\omega_1}_{RH,T} = \theta^{\omega_2}_{RH,T}$. Thus, our problem reduces to a single dimension; and we can consider the function $\zeta(\theta_{RH,T})$ and find its fixed point. More precisely, ζ is a function that gives the optimal threshold if there exists one in the allowed range of θ^{RH}; gives κ^{RH} if the threshold is above the range; and gives 0 if the threshold is below the range. This ensures ζ maps from $[0, \kappa^{RH}]$ to itself. Moreover, if there is a unique equilibrium, the graph of ζ must cross the 45-degree line from above to below.

Claim 18: An equilibrium exists.

An equilibrium will exist given the continuity of ϕ and the fact that Λ is continuous in the threshold sets.

However, the equilibrium reporting distribution is not necessarily unique. Recall that the threshold $\theta_{RH,T}$ is defined as the solution to the equation $u(r_2) - \theta^{RH}\upsilon(\Lambda(r_2)) = u(r_1) - \theta^{RH}\upsilon(\Lambda(r_1))$ or $u(r_2) - u(r_1) = \theta^{RH}(\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1)))$. This describes an individual with $\theta^{LC} = 0$ and a $\theta^{RH} = \bar{\theta}^{RH}$ so that the individual is indifferent between reporting r_1 or r_2. If $\theta^{RH} = 0$ the RHS of this equation is equal to 0. Thus, a sufficient condition for a unique equilibrium is that the RHS is monotonically increasing in $\theta_{RH,T}$ (i.e. the value of θ^{RH} that solves the indifference equation). Unfortunately we cannot guarantee this. As $\theta_{RH,T}$

increases, the probability, conditional on drawing ω_1, of reporting r_1 increases. Similarly, the probability, conditional on drawing ω_2, of reporting r_1 increases. Thus, at r_1 (and similarly r_2) there are both more truth-tellers and more liars, making the change in the difference $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$ ambiguous.

Claim 19: We observe drawing in.

Suppose there is a unique equilibrium and that $f(\omega_2)$ increases. Fixing the input threshold $\theta_{RH,T}$, by Claim 17 the proportion of truth-tellers must increase at r_2. Similarly, the proportion of truth-tellers at r_1 must fall. This makes r_2 relatively more attractive to individuals (compared to r_1). Thus the optimal threshold θ^{RH} (generated by ζ) must rise and we get drawing in.

Claim 20: The model exihibits o-shift and no downwards lying under observability, but may exhibit downwards or not without observability.

Observability will matter as long as some individuals care about the reputation costs. In particular, reputational concerns will imply that individuals would only state the truth or the highest report with observability. We will observe no lying downwards at all under observability of the state by the audience since doing so would incur an LC cost and a reputational cost. Without observability of the state, we may either have lying downwards or not – in the limit if individuals only have LC concerns, then they would never lie down, but in the opposite direction, in the limit if individuals only have reputational concerns then individuals' actions will generically not depend on the drawn state, but only their type, causing lying down..

Claim 21: Depending on parameters, we may observe affinity, aversion or $\hat{g}$-invariance.

Even if the equilibrium of the reporting distribution is unique, we could observe either aversion, affinity or $\hat{g}$-invariance. To see the intuition, note that the $\hat{G}$ treatments do not pin down the new belief about H that subjects hold. Depending on the H, we could get affinity or aversion. In particular, suppose we move from $\hat{G}^A$ (associated with H^A) to $\hat{G}^B$ (where there are two Hs that rationalize $\hat{G}^B$). Imagine that under H^B $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$ increases compared to the difference under H^A, while $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$ decreases under $\tilde{H}^B$compared to H^A. Then we get aversion if subjects believe the new distribution over types is H^B, and we get affinity if subjects believe the new distribution over types is $\tilde{H}^B$.

Formally, we show that two different changes in the exogenous distribution H can both lead to an increase in $\hat{g}(r_2)$. Then we show that they have the opposite implications for $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$. From Figure D.1 we can see that two different shifts of probability mass in H could lead to an increase in $\hat{g^B}(r_2)$ (relative to $\hat{g^A}(r_2)$). The first shifts mass from above $\tau(\omega_1)$ to below it (without altering the relative weights above and below $\tau(\omega_2)$) in Figure D.1. This, fixing the thresholds, doesn't change the reporting of individuals who drew ω_2, but leads to a higher mass of individuals drawing ω_1 to report r_2. This increases $g(r_2)$ but also increases the number of liars at both r_2 and r_1. Recall our fixed point operator that defines the threshold which characterizes the equilibrium: $\zeta(\theta_{RH,T})$. Recall that this, taking as an input everyone else's threshold, returns the optimal threshold. If $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$ increases, this makes the high report less attractive, and so ζ decreases, reducing the equilibrium level of $\theta_{RH,T}$.[49] This reduction will cause aversion. Thus, in order to generate aversion we need that $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$ increases in response to this shift in weight. This can be accomplished simply by ensuring that $\upsilon'(\Lambda(r_2))$ (the derivative of υ) is sufficiently larger than $\upsilon'(\Lambda(r_1))$.

The second shift moves mass from below $\tau(\omega_2)$ to above it (without altering the relative weights above and below $\tau(\omega_1)$). Fixing the thresholds, this doesn't change the reporting of individuals who drew ω_1, but leads to a higher mass of individuals drawing ω_2 to report r_2. This increases $g(r_2)$ but also decreases the number of liars at both r_2 and r_1. If $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$ decreases, this makes the high report more attractive, and so ζ increases. This increases the equilibrium level of $\theta_{RH,T}$, and causes affinity. Similarly to before, in order to generate affinity we need that $\upsilon(\Lambda(r_2)) - \upsilon(\Lambda(r_1))$ decreases in response to this shift in weight. This can be accomplished simply by ensuring that $\upsilon'(\Lambda(r_2))$ is sufficiently larger than $\upsilon'(\Lambda(r_1))$.

Thus, we can get both affinity and aversion (and by continuity $\hat{g}$-invariance) when $\upsilon'(\Lambda(r_2))$ is sufficiently larger than $\upsilon'(\Lambda(r_1))$. Of course, we could get both affinity and aversion but associated with the opposite shifts in weight if we supposed that $\upsilon'(\Lambda(r_2))$ is sufficiently smaller than $\upsilon'(\Lambda(r_1))$. However, since there are more liars at the high report, a sufficiently convex υ will naturally generate the result that $\upsilon'(\Lambda(r_2))$ is sufficiently larger than $\upsilon'(\Lambda(r_1))$, which is what we focus on here. Another sufficient condition is that $\Lambda(r_2)$ responds more to

[49] An equilibrium threshold must fall in this situation. In the case of uniqueness, for any non-trivial parameterization (where at least some types are sometimes willing to lie) we know $\zeta(0) > 0$ (since if no one lies upwards, then it is optimal to best respond by lying upwards). This implies the equilibrium threshold must fall.

the shifts in probability weight than $\Lambda(r_1)$.

We now turn to n states.

Claim 22: Depending on parameters, we may observe drawing in, drawing out or f-invariance.

We provide an example of drawing out (since we have shown drawing in for $n = 2$). Suppose that $n = 3$. Moreover, suppose that individuals only lie one state/report up. Now, move from F^A to F^B by keeping $f^A(\omega_1)$ constant and shifting weight from ω_2 to ω_3. This has two effects. First, fixing strategies, it makes reporting r_3 more attractive (since some of the individuals drawing ω_3 will still report r_3). Second, by the same reasoning, it makes the middle state less attractive. Thus, individuals who draw the lowest state will find reporting the middle state less attractive, and more individuals will simply report the truth. This implies drawing out.

For the Reputation for Honesty + LC model we have ambiguous predictions regarding shifts in $\hat{G}$ even for two states, and this carries over to n states. $\square$

E The Role of Distributional Assumptions

In the body of the paper we suppose that an individual's type is private information, and moreover, the ex-ante prior distribution about types H is non-atomic. In contrast, other papers (M. Dufwenberg and M. A. Dufwenberg 2018, Khalmetski and Sliwka forthcoming, Gneezy et al. 2018) have supposed that there is not necessarily incomplete information about at least some of the dimensions of the type space, and that H has atoms. For example, M. Dufwenberg and M. A. Dufwenberg (2018) consider a model that is related to our Reputation for Honesty model (Appendix C.2), but where everyone has a single known type. Khalmetski and Sliwka (forthcoming) and Gneezy et al. (2018) both consider utility functions that are nested by our Reputation for Honesty + LC model. However, they suppose that there is complete information about the reputational component (although incomplete information about the LC portion of costs).

We made the assumption that H is non-atomic and an individual's type is private information in order to put the models we consider on equal footing, as some models explicitly require a distribution of types and private information about the realized type to generate plausible behavior, e.g., the Reputation for Not Being Greedy model. Recall that our goal of the paper is to understand which types of model can and cannot rationalize the patterns of lying observed in the data.[50] In order to accomplish this, we have attempted to make minimal assumptions on the structure of the utility function. Of course, however, our assumptions regarding private knowledge of types may be substantive, and it is important to understand, in particular, whether it leads us to falsify a class of models which would not be falsified under a different assumption.[51]

It turns out that supposing there is only a single realized type does not change the main finding of our study. The predicted behavior of some models for some of our empirical tests does change if we suppose that H is degenerate and each individual's type is common knowledge, instead of H being non-atomic and the type private knowledge. However, the set of

[50] This is different than the goal of papers whose impetus is to show how much behavior a given model could potentially explain. In this case, making as strong as assumptions as possible, and showing that the behavior one is interested in can still occur, is typically more interesting.

[51] In contrast, this is a lesser problem, given our goal, for those models which cannot be falsified with private knowledge of types. Suppose that, for any of that set of models, common knowledge of types implies the model can be falsified. But, given that the model is not falsified under private information, we should still consider it as a plausible explanation.

falsified models, which we take as our main finding, does not change.[52] First, consider the set of models which we describe as matching Findings 1–4 (listed in Table 2). It turns out that the models that can be falsified by the new tests with binary states when H is non-atomic, can also be falsified when H is degenerate. Six of the nine falsified models listed in Table 2 deliver the exact same prediction for binary states (with the assumption that the G exhibits full support, i.e., we look at full support equilibria). The Reputation for Not Being Greedy model generates different predictions (it now exhibits f-, g- and o-invariance) but is still not in line with the data. The Inequality Aversion + LC and Conformity in LC model can now, depending on parameters, predict drawing in, drawing out, or f-invariance, but otherwise make the same predictions. Thus, supposing that H is degenerate does not lead to different conclusions about how well these models can match the data. The following proposition formalizes this.

Proposition 14 *Suppose $n = 2$. Then all models listed in Table 2, that fail to match the data of our four empirical tests when H is non-atomic and private information, also fail to do so when H is degenerate and common knowledge.*

Proof: For the **LC model**, because an individual engages in a simple one-person optimization problem, the predictions of the model will not change, although all individuals drawing the low state will generically take the same action (since generically individuals will not be indifferent between the two states, and everyone drawing the low state has the same best response). The same reasoning applies to the **Choice Error** model and the **Kőszegi-Rabin + LC** model.

In the **Conformity in LC model**, individuals will never lie down regardless of H. This implies that to observe an equilibrium with full support individuals drawing the low state must weakly prefer to report the low state, i.e., strictly prefer or be indifferent. Thus, we have two cases to consider.

(i) First, suppose the former. If we shift weight in F from the ω_1 to ω_2, with the assumption of a unique equilibrium, we observe f-invariance since no one was willing to lie up before, and the shift in F hasn't increased the aggregate lying costs.

[52] The two models which are not falsified (the Reputation for Honesty + LC model and the LC-Reputation model) also generate different predictions. As explained before, since our goal is to identify models which, under plausible assumptions, fail to match the data, and these models can match the data under some assumptions, we do not focus on them here.

(ii) Next, suppose the latter. Because the equilibrium is unique, there exists a unique proportion of individuals that must be lying up in equilibrium so that individuals drawing the low state are indifferent between reports. This particular proportion doesn't depend on F (it is a feature of the preferences). But, when we shift weight in F from the low to high state, the total proportion of individuals drawing the low state falls. There are two subcases. (a) If after the shift we still observe individuals drawing the low state and reporting the high state, then those drawing the low state must still be indifferent between both report. Then to keep the proportion of individuals lying constant, more individuals drawing the low state need to lie, so we observe drawing in. (b) Alternatively, it could be that after the shift the equilibrium does not feature anyone drawing the low state giving the high report. This would happen if after the shift there are very few individuals who draw the low state, then even if everyone else drawing the low state lies up, it is not a best reponse for someone drawing the low state to give the high report (recall that lying costs are normalized by the average amount of lying). Thus, since the equilibrium features no individuals drawing the low state and giving the high report, we have drawing out. We observe affinity, o-invariance, and no lying down for the same reasons as in the body of the paper.

We next consider **Inequality Aversion**. Because individuals' utility does not depend on their drawn state, to get full support it must be the case that all individuals are indifferent between the two states. However, the set of equilibria will not vary with F, for the same reason as in the body of the paper. The rest of the results do not change.

In the **Inequality Aversion + LC model** there are several possibilities.

(i) First, individuals drawing each state could strictly prefer to report their state (because of the LC cost, it can never be the case that those drawing the low state strictly prefer to report high and vice versa). In this case, increases in$f(\omega_2)$ will increase the fraction of individuals reporting r_2, making the high state more attractive relative to the low state, and so cause either f-invariance or drawing in.

(ii) The second possibility is that those drawing the high state strictly prefer to give the high report and those drawing the low state are indifferent. There are three subcases. (a) If after the increase in $f(\omega_2)$ individuals drawing the low state are still indifferent in equilibrium, the probability of reporting high, conditional on drawing the low state, must have fallen. This implies we observe drawing out. (b) If we moved to an equilibrium without full support we could have drawing in, since after the shift, there are no longer enough individuals drawing the low state and reporting the low state to maintain indifference. (c) The third case is that, after the shift in F, those drawing the low state now strictly prefer to give the low report and those drawing the high state are indifferent. This can generate either drawing in or drawing out. The former could occur because individuals who draw the high state now are a high enough fraction so that, if none of them lie down, they all prefer to give the high report. The latter could occur because to maintain indifference between the two reports, the probability of reporting low, conditional on drawing high, must increase. Thus, depending on parameters, we can have drawing in, drawing out or f-invariance. We observe affinity, o-invariance, and, depending on parameters, lying down or not for the same reasons as in the body of the paper.

In the **Censored Conformity in LC model**, individuals will never lie down regardless of H. This implies that, to observe an equilibrium with full support, individuals drawing the low state must weakly prefer to report the low state, i.e., strictly prefer or be indifferent. We consider each case separately.

(i) In the former case, as in the Conformity in LC model described above, we will observe f-invariance.

(ii) In the latter case, there is a unique proportion, conditional on drawing the low state, that must report the high state, in order to ensure that individuals drawing the low state are indifferent. This proportion doesn't change with F. Recall that in the Censored Conformity in LC model the LC costs are "normalized" by the average lying cost among those who could lie, which is the average lying cost of those who drew the low state, or the proportion of those drawing the low state and reporting the high state. Since, as just described, the equilibrium value of this doesn't change with F, we still observe f-invariance. We observe affinity, o-invariance, and no lying down for the same reasons as in the body of the paper.

In the **Reputation for Not Being Greedy model**, individuals care about their monetary payoff and their estimated type. If individuals' types are known then the second motivation disappears, and individuals behave exactly as if they simply want to maximize their monetary payoff; and so will exhibit f, $\hat{g}$ and o-invariance and no lying down.

We next consider **Guilt Aversion**. Because individuals' utility does not depend on their drawn state, to get full support it must be the case that all individuals are indifferent between the two states. However, the set of equilibria will not vary with F, for the same reason as in the body of the paper. Shifts in $\hat{G}$ also induce the same effects, observability does not change behavior, and we wil observe lying down for the same reasons also. □

Second, consider the set of models that, given our assumption on H, fail to match Findings 1–4 (discussed in Appendix C). These consist of the standard model, the Reputation for Honesty model and the Audit model. As should be relatively clear from the previous discussions, the standard model's predictions do not depend on our assumptions regarding H and the Audit model still fails to match the stylized findings, for the same reason as when H is non-atomic. However, the predictions of the Reputation for Honesty model with a degenerate H differ from the predictions in Appendix C. A degenerate H implies individuals must be indifferent between all reports that are made with positive probability in equilibrium. Since individuals can randomize differently based on their drawn state, equilibria can be constructed that have full support and thus Reputation for Honesty with degenerate H can explain Findings 1–4. However, such a model fails to match the data from our new tests, in particular the $\hat{G}$ treatments.

Proposition 15 *Suppose subjects' utility functions are as in the Reputation for Honesty model but H is degenerate and common knowledge. Then, for $n = 2$, we have affinity.*

Proof: A degenerate H implies individuals must be indifferent between all reports that are made with positive probability in equilibrium (since if one subject had a strict preference for one report, all subjects would exhibit the same strict preference). Given indifference, subjects can randomize differently based on their drawn state. In the $\hat{G}$ treatments, $\hat{G}$ cannot provide information about H since this is already common knowledge. It can only provide information about which equilibrium (out of the multiple potential equilibria) is being selected. The treatments induce a belief $\hat{G}$ about the equilibrium distribution of reports, and thus subjects' equilibrium strategy generates a reporting distribution $G = \hat{G}$.[53] Thus, if a "higher" $\hat{G}$ (in the sense of representing a higher average report) is induced, then a "higher" G will result. This implies affinity. □

[53] If we only assume best-response behavior, then any behavior in the $\hat{G}$ treatments can be rationalized. This is because all subjects play a mixed strategy and are thus indifferent between the different reports. However, in order to support $\hat{G}$ as an equilibrium distribution, it has to be the case that subjects play $G = \hat{G}$ to preserve the indifference of the other players.

The prediction of affinity is not in line with the data from our $\hat{G}$ treatments.

Third, we can use a particular aspect of the OBSERVABLE treatment to further distinguish between models. In the OBSERVABLE treatment, we know the true state ω of subjects. We find that subjects who drew the same state differ in their behavior. Some report honestly ($r = \omega$) and others lie up ($r > \omega$) (see Figure 7). Such within-state heterogeneity can be generated, in a robust way (in the sense explained below), by models with non-atomic H (our maintained assumption outside this appendix) and that is a reason why we do not focus on this behavioral regularity in the body of the text. In particular, it is straightforward to show that this pattern of behavior can be robustly generated by the two models that our empirical exercise cannot falsify, Reputation for Honesty + LC and LC-Reputation. However, this behavior is at odds with several of our models if we assume a degenerate H. In particular, as the next proposition shows, this behavior cannot be generated in a way that is robust to perturbations in θ. It can only occur for an isolated set of points in at least one of the dimensions of Θ. In other words, suppose we begin with a situation where individuals drawing the same state make different reports – if we peturbed individuals' common θs then all individuals drawing the same state would make the same report.

Proposition 16 *Suppose H is degenerate and the drawn state is observed by the audience as in our OBSERVABLE treatment. Then under the LC, Reputation for Honesty+LC, LC-Reputation, Reputation for Being Not Greedy, Reputation for Honesty and Audit models we observe individuals drawing the same state and making the same report only for a discrete subset of at least one dimension of Θ.*

Proof: Assume subjects have LC utility. Then r and r' are both reported if and only if $\phi(r, c(r, \omega; \theta^{LC}) = \phi(r', c(r', \omega; \theta^{LC})$. Observe that, for any $\theta^{LC'}$ in a neighborhood around θ^{LC}, by the assumptions on cross partials $\phi(r, c(r, \omega; \theta^{LC'}) \neq \phi(r', c(r', \omega; \theta^{LC'})$. Moreover, we can always find a small enough neighborhood such that for all $\theta^{LC'}$ no other indifferences occur. This shows that that if we peturb θ^{LC} we break indifference and so any θ^{LC} generating indifference must be isolated. The result follows by the definition of a discrete set.
The Reputation for Honesty+LC model reduces to the LC model plus an additional fixed cost of lying if states are observed, and the previous result thus carries over. The LC-Reputation

model reduces to (a monotone transformation of) the LC model if θ^{LC} is known, i.e., the same result obtains. Under the Reputation for Being Not Greedy model, if θ^{RNG} is known, then the model reduces to (a monotone transformation of) the standard model. This means the result obtains (since we know the standard model generates a degenerate G).
The Reputation for Honesty model, under observability, reduces to an LC model with a fixed cost of lying. The fixed cost is the same for all individuals with a degenerate H and so the result above follows. The Audit model under observability reduces to an LC model with zero cost of lying down and a fixed cost of lying up; thus the LC model result follows. □

In contrast, the other models we consider in our paper can generate within-state heterogeneity in the OBSERVABLE treatment robustly even if H is degenerate. The Choice Error model generates a distribution of reports for any given single $\theta^{CE} < \infty$. The Conformity in LC, Censored Conformity in LC, Inequality Aversion, Inequality Aversion+LC and Guilt Aversion models still feature non-trivial equilibrium considerations and thus allow for mixing across reports. For example, consider the Conformity in LC model with $n = 2$. It could be the case that given a particular θ^{CLC}, we observe individuals drawing the low state giving both the low and high report. Fixing others' behavior, adusting the preference parameter slightly will break indifference. But equilibrium behavior can adjust to maintain overall indifference. Suppose, for example, θ^{CLC} increases slightly. Then more individuals could lie up and under the new equilibrium indifference between making the low and high report could be maintained.[54] This can also occur in the Kőszegi-Rabin + LC model when a PPE may involve randomization; the adjustments are made not to equilibrium strategies of other players as in the Conformity in LC model, but rather by the individual themselves.

[54] This behavior is linked to the fact that there are multiple equilibria.

F Additional Experiments

In this Appendix we present two additional sets of experiments that we conducted to test specific predictions of some of the models considered in the paper.

Our first set of additional experiments test predictions of the LC model regarding specific shifts in the distribution F for n states. We can show that if we change the distribution of F, but only for the highest M states, then the LC models predicts that the distribution of reports will not change for the lowest $n - M$ states. Essentially, changes in F for the highest states do not cause changes in G for lower states/reports.

Proposition 17 *Under LC, consider two distributions F^A and F^B such that $f^A(\hat{\omega}) = f^B(\hat{\omega})$ for all $\hat{\omega} \leq \omega^*$. Then for all $\hat{r} \leq r^* = \omega^*$: $g^A(\hat{r}) = g^B(\hat{r})$.*

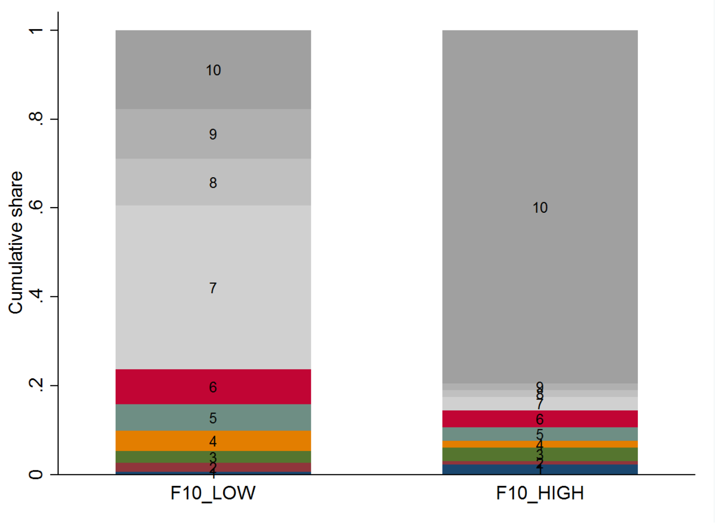
Proof: Recall no individuals lie down in the LC model. Moreover, the optimal report by an individual is a function only of θ^{LC} and of ω. Thus, conditional on drawing an $\omega \leq \omega^*$, any decision-maker's best response is the same under F^A and F^B (for a given θ^{LC}). Thus, the distribution of reports for $\hat{r} \leq r^* = \omega^*$ must be the same. $\square$

To test this prediction, we use an experiment with 10-state distributions. The setup is identical to that described in the main paper 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. The LC model predicts that there will be no difference in the fraction of subjects reporting numbers 1–6. These experiments were conducted in Nottingham between May and June 2015 with a total of 284 subjects.

We find a significant difference in the distribution of reports of our F10 treatments. Figure F.1 shows the distribution of reports across the two treatments. Fewer subjects report 1 to 6 in F10_HIGH than F10_LOW (14 percent vs. 24 percent, p= 0.045, OLS with robust SE; p = 0.048, χ^2 test). Thus, shifting the probability of high outcomes in the right tail of the distribution draws in subjects from the left tail of the distribution.

This finding is not in line with the predictions of the LC model. The concurrent papers by Gneezy et al. (2018) and Garbarino et al. (forthcoming) also 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.

Figure F.1: Distribution of reports in F10_LOW and F10_HIGH



The second set of additional experiments tests some specific predictions of the Kőszegi-Rabin + LC model regarding the role of expectations using a design that follows closely the design of Abeler et al. (2011). Subjects report ten times the outcome of a coin flip. Their earnings are equal to the number of tails they report in pounds. However, subjects' reports are only paid out with 50 percent probability, and with the other 50 percent subjects receive a fixed payment which differed by treatment. In one treatment (KR_HIGH) the fixed payment is £8, while in the other (KR_LOW) it is £4. The payment lottery is only resolved after subjects made their report. Because the fixed payment enters expectations, the Kőszegi-Rabin + LC model predicts that subjects will lie more if the fixed payment is higher. These experiments were conducted in Oxford in October 2013 with a total of 155 subjects.

We find no significant difference between treatments. The average report is 6.49 in

KR_HIGH and 6.36 in KR_LOW, the difference is not statistically significant (p=0.676, OLS with robust SE; p = 0.651, Wilcoxon rank-sum test).

G 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.

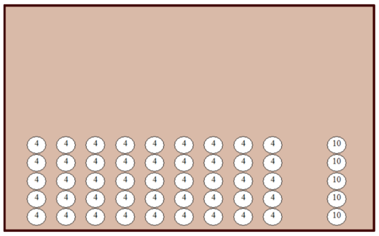
G.1 Instructions for F_LOW

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.

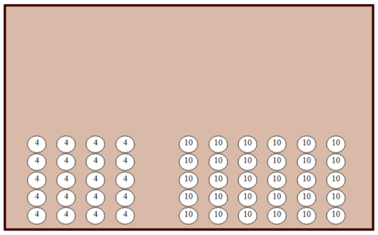
The on-screen instructions about how to perform the draw were as follows:
When you have finished your questionnaire click the CONTINUE button to proceed to the next screen where you will find instructions for how to perform the draw that will determine your payment.

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.

G.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.

G.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.

G.4 Instructions for G_LOW

Welcome and thank you for taking part in this study.

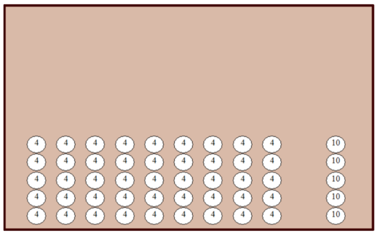
The study is run by the “Centre for Decision Research and Experimental Economics” and has been financed by various research foundations. During the study, we request that you turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.

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

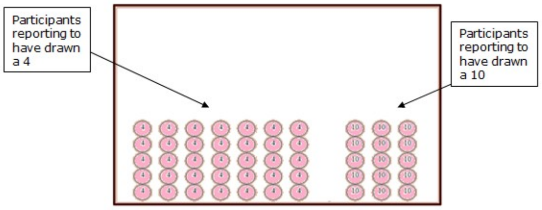
Imagine that we would run the following experiment here in this lab with 50 participants. Participants are invited by email and asked to complete a questionnaire about their money management. Participants receive an amount of money for completing the questionnaire. To determine how much money they are 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.


Participants reporting to have drawn a 4
Participants reporting to have drawn a 10
4
10

[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:______________________

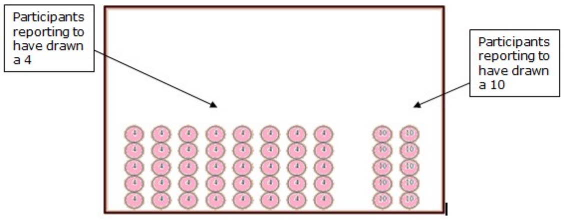
4) Some of the participants who drew a 4 actually reported 4. Can you imagine why they would do that? Your answer:______________________

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.


Participants reporting to have drawn a 4
Participants reporting to have drawn a 10

[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) ____________________________________

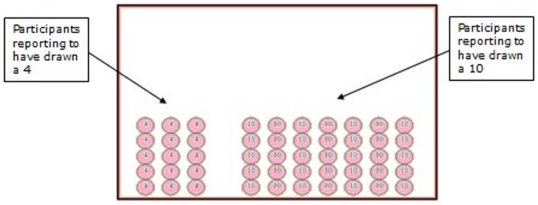
YOUR PAYMENT FOR TAKING PART IN TODAY'S STUDY

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.

G.5 Instructions for G_HIGH

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

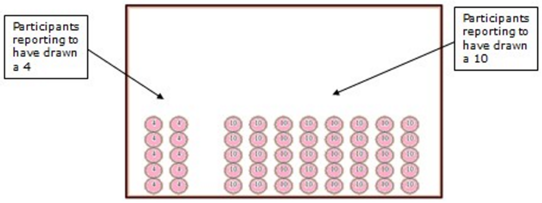
[3] Of the 50 participants, 15 report to have drawn a 4 and 35 report to have drawn a 10. This outcome is shown in the figure below.


Participants reporting to have drawn a 4
Participants reporting to have drawn a 10

[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.


Participants reporting to have drawn a 4
Participants reporting to have drawn a 10

[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?

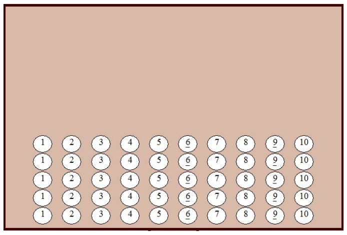
[9] 10) Which of the two imagined outcomes described above do you think is more realistic? Your answer: The outcome where 35 out of 50 participants reported a 10; The outcome where 40 out of 50 participants reported a 10

G.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.

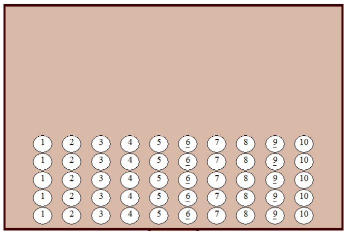
Then bring the questionnaire and the payment sheet to the experimenter who is waiting outside.

G.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.

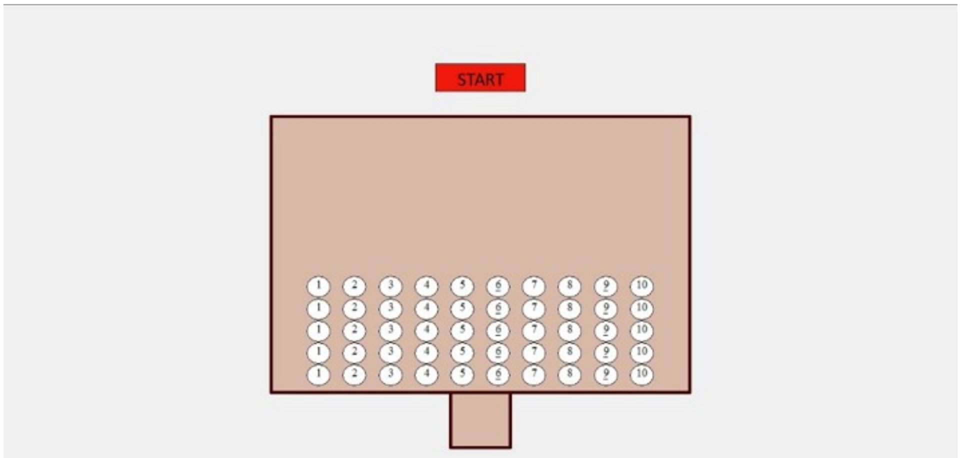
After the draw, open the brown envelope that is placed on your desk. The envelope contains 10 coins of £1 each. Take as many coins as the number of the chip you have drawn. Then turn off your computer using the power button and quietly exit the lab leaving these instructions, your questionnaire, and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

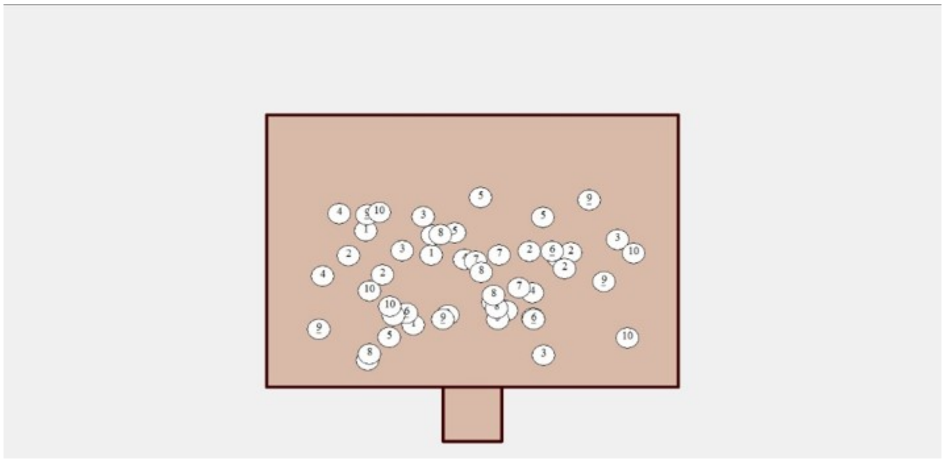
If you have a question now, please raise your hand and the experimenter will come to your desk to answer it.

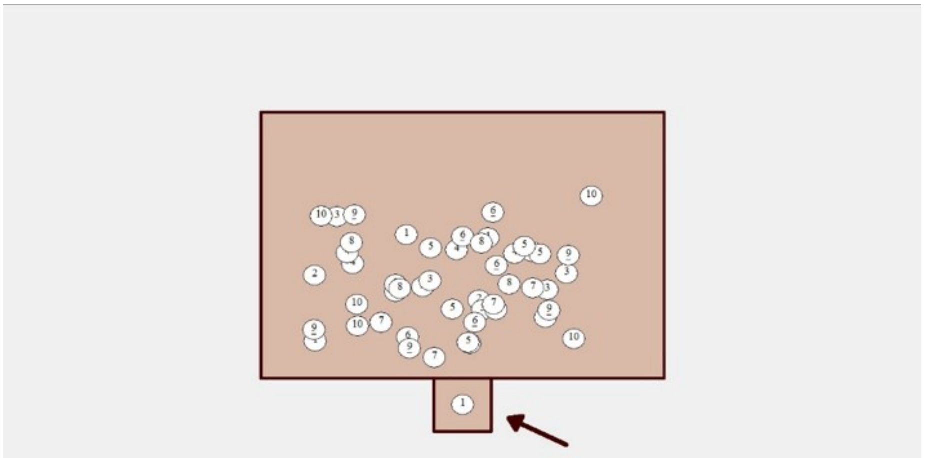
The on-screen instructions about how to perform the draw were as follows:

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.



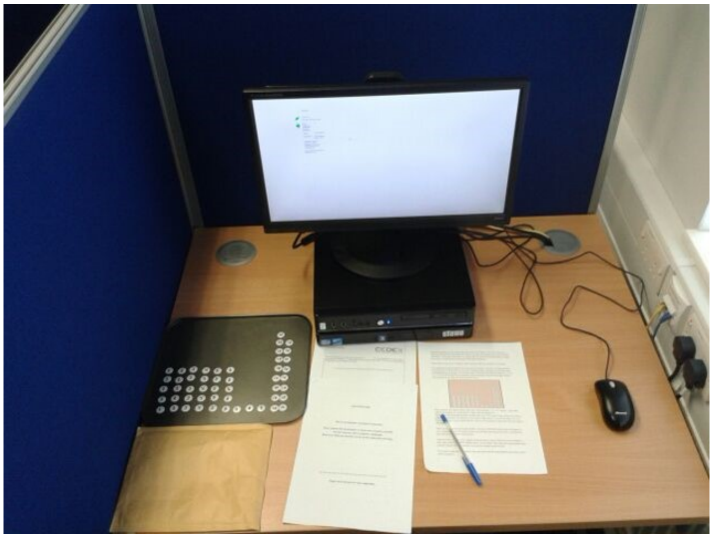




Please now open the brown envelope that you can find on your desk. The envelope contains 10 coins of £1 each. Take as many coins as the number of the chip you have drawn.

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.)

G.8 Laboratory Setup



H Calibration Details

H.1 Details of the Conformity in LC Calibration

This section describes the details of the calibration of the Conformity in LC model presented in Section 4. We calibrate the Conformity in LC 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 $r - \theta^{CLC}\frac{c(r,\omega)}{\bar{c}}$. $c(r,\omega)$ 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 Conformity in LC model, this simply represents the fraction of people who drew ω_1 but report r_2. We normalize $r_1 = -1$ and $r_2 = 1$ in line with our normalized payoffs in the meta study. Moreover, we will suppose that θ^{CLC} is uniformly distributed on $[0, \kappa^{CLC}]$. Given an equilibrium with full support the threshold type (who draws the low state) must satisfy the condition $1 - \bar{\theta}^{CLC}\frac{1}{\bar{c}} = -1$ or $\bar{\theta}^{CLC} = 2\bar{c}$. 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. Thus $\bar{\theta}^{CLC} = 0.7$. Moreover, the fraction of liars, conditional on drawing the low state (which in the F_LOW treatment happened with probability equal to 0.9), is equal to $\frac{\bar{\theta}^{CLC}}{\kappa^{CLC}} = \frac{.7}{\kappa^{CLC}} = \frac{0.35}{0.9}$. In other words $\kappa^{CLC} = 1.8$. Given this, suppose that $f(\omega_1) = 0.9$ 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.

H.2 Details of the Reputation for Honesty + LC Calibration

This section describes the details of the calibration of the Reputation for Honesty + LC model presented in Section 4. When there are six states, observe that because the fixed cost is 3, individuals who draw ω_3 and above will never want to lie. Moreover, individuals who draw ω_1 will never lie to below r_5 and individuals who draw ω_2 will only lie to r_6. This immediately implies there are no liars at r_1, r_2, r_3 and r_4. In constructing the equilibrium we suppose that there are some individuals who drew ω_1 who want to report r_5, and some who want to report r_6. Similarly, we suppose there are some individuals who are willing to lie to r_6 conditional on drawing ω_2. We then verify this is the case.

The threshold type, conditional on drawing ω_1, that is indifferent between reporting r_1 and r_5 is defined by $\theta^1_{1,5} = \frac{4-c}{\Lambda(r_5)}$. The threshold, conditional on drawing ω_1, between reporting r_5 and r_6 is $\theta^1_{5,6} = \frac{1-c}{\Lambda(r_6)-\Lambda(r_5)}$. Similarly, the threshold type, conditional on drawing ω_2, that is indifferent between reporting r_2 and r_6 is $\theta^2_{2,6} = \frac{4-c}{\Lambda(r_6)}$. Using these thresholds, we find that $\Lambda(r_5) = \frac{\frac{1}{6k}(\theta^1_{1,5}-\theta^1_{5,6})}{\frac{1}{6k}(\theta^1_{1,5}-\theta^1_{5,6})+\frac{1}{6}}$ and $\Lambda(r_6) = \frac{\frac{1}{6k}(\theta^1_{5,6}+\theta^2_{2,6})}{\frac{1}{6k}(\theta^1_{5,6}+\theta^2_{2,6})+\frac{1}{6}}$. We then find the fixed point, i.e., the equilibrium. In addition to the values highlighted in the text, it is also the case that $\theta^1_{1,5} \approx 6.67$, $\theta^1_{5,6} \approx 4.5$ and $\theta^2_{2,6} \approx 2.7$. We thus verify our assumptions on the structure of the equilibrium made in the previous paragraph (i.e., the thresholds are in line with our assumptions).

In the case with two states (remember that they pay 1 and 6), and no lying down in equilibrium, we have a single threshold type for those drawing the low state $\theta^1 = \frac{5-c}{\Lambda(r_6)}$. The fraction of liars at the high report is $\Lambda(r_6) = \frac{f(\omega_1)\frac{1}{k}\theta^1}{f(\omega_1)\frac{1}{k}\theta^1+(1-f(\omega_1))}$. For $f(\omega_1) = 0.4$ we find $\theta^1 \approx 7.1$, and for $f(\omega_1) = 0.9$ we find $\theta^1 \approx 2.9$.

When we allow for lying down in equilibrium we now have two thresholds, one for each state: $\theta^1 = \frac{5-c}{\Lambda(r_6)-\Lambda(r_1)}$ and $\theta^6 = \frac{5+c}{\Lambda(r_6)-\Lambda(r_1)}$. The fraction of liars at each report is $\Lambda(r_6) = \frac{f(\omega_1)\frac{1}{k}\theta^1}{f(\omega_1)\frac{1}{k}\theta^1+(1-f(\omega_1))\frac{1}{k}\theta^6}$ and $\Lambda(r_1) = \frac{(1-f(\omega_1))\frac{1}{k}(k-\theta^6)}{f(\omega_1)\frac{1}{k}(k-\theta^1)+(1-f(\omega_1))\frac{1}{k}(k-\theta^6)}$. No equilibrium of this type exists when $f(\omega_1) = 0.4$, but when $f(\omega_1) = 0.9$ we find $\theta^1 \approx 2.5$ and $\theta^2 \approx 9.9$.

We last show that given our calibration, for the equilibrium induced by $f = 0.9$ that features lying down, the derivative of $\Lambda(r_6)$ is larger than the derivative of $\Lambda(r_1)$ with respect to the shifts in H that $\hat{G}$ could induce. As shown in the proof for the Reputation for Honesty + LC model in Proposition 2, this implies shifting probability mass of H from above θ^1 (but not above θ^6) to below it will cause aversion, and shifting weight from below θ^6 (but above θ^1) to above it will cause affinity. However, both will cause an increase in $\hat{g}(r_6)$. Simple calculation indeed verifies that for both shifts in weight (in H) the derivative of $\Lambda(r_6)$ is larger than the derivative of $\Lambda(r_1)$.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman (2011). “Reference points and effort provision”. *American Economic Review* 101.2, pp. 470–492.

Andreoni, James and Douglas Bernheim (2009). “Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects”. *Econometrica* 77.5, pp. 1607–1636.

Batson, Daniel, Diane Kobrynowicz, Jessica Dinnerstein, Hannah Kampf, and Angela Wilson (1997). “In a very different voice: Unmasking moral hypocrisy.” *Journal of Personality and Social Psychology* 72.6, p. 1335.

Batson, Daniel, Elizabeth Thompson, Greg Seuferling, Heather Whitney, and Jon Strongman (1999). “Moral hypocrisy: Appearing moral to oneself without being so”. *Journal of Personality and Social Psychology* 77.3, p. 525.

Battigalli, Pierpaolo and Martin Dufwenberg (2007). “Guilt in games”. *American Economic Review* 97.2, pp. 170–176.

— (2009). “Dynamic psychological games”. *Journal of Economic Theory* 144.1, pp. 1–35.

Bénabou, Roland and Jean Tirole (2006). “Incentives and Prosocial Behavior”. *American Economic Review* 96.5, pp. 1652–1678.

Bolton, Gary and Axel Ockenfels (2000). “ERC: A theory of equity, reciprocity, and competition”. *American Economic Review* 90.1, pp. 166–193.

Charness, Gary and Martin Dufwenberg (2006). “Promises and partnership”. *Econometrica* 74.6, pp. 1579–1601.

Cojoc, Doru and Adrian Stoian (2014). “Dishonesty and charitable behavior”. *Experimental Economics* 17.4, pp. 717–732.

Conrads, Julian, Bernd Irlenbusch, Rainer Michael Rilke, Anne Schielke, and Gari Walkowitz (2014). “Honesty in tournaments”. *Economics Letters* 123.1, pp. 90–93.

d’Adda, Giovanna, Donja Darai, Nicola Pavanini, and Roberto A Weber (2017). “Do leaders affect ethical conduct?” *Journal of the European Economic Association* 15.6, pp. 1177–1213.

Demichelis, Stefano and Jörgen W Weibull (2008). “Language, meaning, and games: A model of communication, coordination, and evolution”. *American Economic Review* 98.4, pp. 1292–1311.

Diekmann, Andreas, Wojtek Przepiorka, and Heiko Rauhut (2015). “Lifting the veil of ignorance: An experiment on the contagiousness of norm violations”. *Rationality and Society* 27.3, pp. 309–333.

Dufwenberg, Martin and Martin A. Dufwenberg (2018). “Lies in Disguise – A Theoretical Analysis of Cheating”. *Journal of Economic Theory* 175, pp. 248–264.

Ellingsen, Tore and Magnus Johannesson (2008). “Pride and prejudice: The human side of incentive theory”. *American Economic Review* 98.3, pp. 990–1008.

Ellingsen, Tore and Robert Östling (2010). “When does communication improve coordination?” *American Economic Review* 100.4, pp. 1695–1724.

Fehr, Ernst and Klaus M Schmidt (1999). “A theory of fairness, competition, and cooperation”. *Quarterly Journal of Economics* 114, pp. 817–868.

Fischbacher, Urs and Franziska Föllmi-Heusi (2013). “Lies in disguise—an experimental study on cheating”. *Journal of the European Economic Association* 11.3, pp. 525–547.

Frankel, Alex and Narvin Kartik (forthcoming). “Muddled Information”. *Journal of Political Economy.*

Gächter, Simon and Jonathan F Schulz (2016). “Intrinsic honesty and the prevalence of rule violations across societies”. *Nature* 531, pp. 496–499.

Garbarino, Ellen, Robert Slonim, and Marie Claire Villeval (forthcoming). “Loss Aversion and lying behavior: Theory, estimation and empirical evidence”. *Journal of Economic Behavior and Organization.*

Gibson, Rajna, Carmen Tanner, and Alexander Wagner (2013). “Preferences for truthfulness: Heterogeneity among and within individuals”. *American Economic Review* 103, pp. 532–548.

Gneezy, Uri (2005). “Deception: The role of consequences”. *American Economic Review* 95.1, pp. 384–394.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel (2018). “Lying Aversion and the Size of the Lie”. *American Economic Review* 108.2, pp. 419–453.

Gneezy, Uri, Bettina Rockenbach, and Marta Serra-Garcia (2013). “Measuring lying aversion”. *Journal of Economic Behavior & Organization* 93, pp. 293–300.

Greene, Joshua and Joseph Paxton (2009). “Patterns of neural activity associated with honest and dishonest moral decisions”. *Proceedings of the National Academy of Sciences* 106.30, pp. 12506–12511.

Grossman, Zachary (2015). “Self-signaling and social-signaling in giving”. *Journal of Economic Behavior & Organization* 117, pp. 26–39.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter (2008). “Antisocial punishment across societies”. *Science* 319.5868, pp. 1362–1367.

Hugh-Jones, David (2016). “Honesty, beliefs about honesty, and economic growth in 15 countries”. *Journal of Economic Behavior & Organization* 127, pp. 99–114.

Hurkens, Sjaak and Navin Kartik (2009). “Would I lie to you? On social preferences and lying aversion”. *Experimental Economics* 12.2, pp. 180–192.

Jiang, Ting (2013). “Cheating in mind games: The subtlety of rules matters”. *Journal of Economic Behavior & Organization* 93, pp. 328–336.

Kajackaite, Agne and Uri Gneezy (2017). “Incentives and cheating”. *Games and Economic Behavior* 102, pp. 433–444.

Kartik, Navin, Olivier Tercieux, and Richard Holden (2014). “Simple mechanisms and preferences for honesty”. *Games and Economic Behavior* 83, pp. 284–290.

Khalmetski, Kiryl and Dirk Sliwka (forthcoming). “Disguising Lies – Image Concerns and Partial Lying in Cheating Games”. *American Economic Journal: Microeconomics*.

Kőszegi, Botond and Matthew Rabin (2006). “A model of reference-dependent preferences”. *Quarterly Journal of Economics* 121.4, pp. 1133–1165.

— (2007). “Reference-dependent risk attitudes”. *American Economic Review* 97.4, pp. 1047–1073.

Levine, David (1998). “Modeling altruism and spitefulness in experiments”. *Review of Economic Dynamics* 1.3, pp. 593–622.

Luce, R Duncan (1959). *Individual choice behavior*. John Wiley.

Mann, Heather, Ximena Garcia-Rada, Lars Hornuf, Juan Tafurt, and Dan Ariely (2016). “Cut from the Same Cloth: Surprisingly Honest Individuals Across Cultures”. *Journal of Cross-Cultural Psychology* 47.6, pp. 858–874.

Mazar, Nina, On Amir, and Dan Ariely (2008). “The dishonesty of honest people: A theory of self-concept maintenance”. *Journal of Marketing Research* 45.6, pp. 633–644.

McFadden, Daniel et al. (1973). “Conditional logit analysis of qualitative choice behavior”.

Pascual-Ezama, David et al. (2015). “Context-dependent cheating: Experimental evidence from 16 countries”. *Journal of Economic Behavior & Organization* 116, pp. 379–386.

Rauhut, Heiko (2013). “Beliefs about Lying and Spreading of Dishonesty: Undetected Lies and Their Constructive and Destructive Social Dynamics in Dice Experiments”. *PLoS One* 8.11.

Ruedy, Nicole and Maurice Schweitzer (2010). “In the moment: The effect of mindfulness on ethical decision making”. *Journal of Business Ethics* 95.1, pp. 73–87.

Shalvi, Shaul, Jason Dana, Michel Handgraaf, and Carsten De Dreu (2011). “Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior”. *Organizational Behavior and Human Decision Processes* 115.2, pp. 181–190.

Tadelis, Steven (2011). “The power of shame and the rationality of trust”. *Haas School of Business Working Paper*.

Townsend, Robert (1979). “Optimal contracts and competitive markets with costly state verification”. *Journal of Economic Theory* 21.2, pp. 265–293.