Signaling Motives in Lying Games

Tilman Fries (WZB Berlin)

Discussion Paper No. 269

January 14, 2021

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

Signaling motives in lying games

Tilman Fries*

January 12, 2021

Abstract

This paper studies the implications of agents signaling their moral type in a lying game. In the theoretical analysis, a signaling motive emerges where agents dislike being suspected of lying and where some types of liars are more stigmatized than others. The equilibrium prediction of the model can explain experimental data from previous studies, in particular on partial lying, where some agents dishonestly report a non payoff-maximizing report. I discuss the relationship with previous theoretical models of lying that conceptualize the image concern as an aversion to being suspected of lying. The second half of the paper tests the theoretical predictions in an experiment. In contrast to previous literature, the experimental results show no evidence that image concerns influence lying behavior in the laboratory.

Keywords: lying, image concerns, honesty, experiment
JEL Codes: C91, D82, D90

*WZB Berlin Social Science Center, Reichpietschufer 50, D-10785 Berlin; e-mail: tilman.fries@wzb.eu. I am grateful to Johannes Abeler, Kai Barron, Christian Basteck, Martin Dufwenberg, Dirk Engelmann, Hoa Ho, Agne Kajackaite, and Daniel Parra for helpful comments and discussions. I further thank participants at the ESA World Meetings 2020 and participants at the seventh CRC 190 Retreat. Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged.

1 Introduction

The virtue ethics of the ancient Greeks recognize honesty among the desirable moral characteristics which can lead individuals to flourish and to live a "good life".[1] In contemporary popular science, intellectuals stress the value of honesty.[2] Virtues also play a role in economic situations; if Alice is a buyer and Bob is a seller in a credence good market, it will be relevant for Alice not just to ask if Bob was honest with her in the exchange they just had, but whether Bob will be honest again in future exchanges. To form this latter expectation, Alice needs to have an idea about Bob's moral character, in particular about his honesty. This paper is concerned with the strategic implications when individuals want to appear honest.

In strategic situations where some agents are better informed than others, truthful communication can be difficult or impossible. This impedes information transmission and can lead to market failures (Akerlof, 1970, Crawford and Sobel, 1982). Some of these inefficiencies can be overcome if lying is costly for agents (Kartik, 2009), but the size and form of lying costs is mainly an empirical question.

More recently, a literature has emerged that empirically investigates lying costs in laboratory experiments. In an experiment, Fischbacher and Föllmi-Heusi (2013)–or F&FH–gave participants a six-sided die. Participants were instructed to roll the die in private and report the number they rolled to the experimenter. Upon reporting, participants received a payoff in Swiss Franks that corresponded to their die roll, with the exception of the number six, which paid nothing. Since the objective distribution of the die roll is known, lying behavior can be inferred from the aggregate report distribution. F&FH find that the empirical distribution of reports is consistent with some participants reporting honestly and other participants lying. In various follow-up experiments–that sometimes let participants flip coins instead of rolling a die–similar patterns emerge (Abeler, Nosenzo, and Raymond, 2019).

One robust feature in experiments that use the F&FH die-roll task is that some individuals lie and dishonestly report a four when they could have earned more money by lying and reporting a five. One reason for the observed behavior could be that individuals dislike being suspected of lying; since fewer individuals lie to report a number that does not maximize their monetary payoff, reporting a lower number evokes less suspicion. Papers by Dufwenberg and Dufwenberg (2018), Gneezy, Kajackaite, and Sobel (2018) and Khalmetski and Sliwka (2019) provide theoretical models that formalize this intuition.[3] In doing so, they all have to come to terms with the fact that lying decisions depend on perceived suspicion which in turn depends on lying decisions. Suspicion therefore is an equilibrium outcome of a game between an agent

[1] See e.g. the Stanford Encyclopedia article on Virtue Ethics (Hursthouse and Pettigrove, 2018).

[2] For example, in Jordan Peterson's book "12 Rules for Life" (Peterson, 2018), which enjoyed considerable media attention and commercial success, one rule is "Tell the truth – or, at least, don't lie". Another example is that of Adam Grant, a psychologist and author who, on Sep. 7 2019 sent out the following tweet to his twitter followers, which gathered more than 900 retweets and 2800 likes: *"When you face a choice between being polite and being honest, err on the side of the truth. It's better to be disliked but respected than to be liked but disrespected. In the long run, the people we trust the most are those who have the courage to be sincere. #SaturdayThoughts"* (Grant, 2019)

[3] From now on in the text I will refer to them as D&D, GK&S, and K&S respectively.

and an observer, in which an agent draws a state (a number on a die, a coin flip) and makes a report to an observer. The report serves as a signal to the observer, who in turn forms a belief about the likelihood that the agent lied; a measure of suspicion. The agent's utility is *belief-dependent*, as it depends on the *image* that the observer attaches to the agent after hearing their report. In their meta-study, Abeler et al. (2019)–from now on AN&R–conclude that such image concerns are key to explain the stylized empirical facts observed in experiments on lying.

While image concerns are deemed to be important, there are different ways to conceptualize them. AN&R find that two kinds of image concerns can explain the observed empirical regularities in lying games. The first is an image concern that (in various forms) is used in models by D&D, GK&S and K&S, where individuals want to signal that they did not lie.[4] The second is a lying model where the signaling motive is similar to the prosocial image model of Bénabou and Tirole (2006)–hereafter B&T. In this model, individuals want to appear as someone who is intrinsically honest. The main difference between both approaches is that in the former individuals want to signal a good deed (they did not lie), whereas in the latter model individuals want to signal a moral character (they are intrinsically honest). In this paper I ask if this second approach to image concerns can provide useful insights and extend our understanding of lying behavior. I derive a lying model based on B&T, which so far has only received cursory attention in the literature.[5]

I study the strategic implications of individuals signaling their moral character in a lying game. Agents draw a random number (by rolling a die, flipping a coin, etc.) and make a report to an observer. They are morally concerned and incur a cost if their report does not equal their draw. Agents differ in the extent to which they are morally concerned; some suffer high and others low costs from lying. Individual types are private, but in equilibrium the agents' reports are informative about their type. This happens because worse moral types have lower lying costs, and are therefore more likely to dishonestly report a high number, than better types. In the model, *credibility* and *quality* of the report influences an agent's social image. A report is more *credible* the more likely it is that it was made truthfully. Moreover, the reputation attached to a report depends on the moral type, or *quality*, of the liars reporting it.

To illustrate how reputations form in the model, consider the following example of a professor who, at the day of a final exam, receives messages from some of her students that they are sick and cannot participate in the exam. By university guidelines, sickness is the only acceptable excuse for not writing the exam. There might however be other reasons that induce a student to write that they are sick. Specifically, any student might be *sick* with some probability, or, in case they are not sick, they might be in an *emergency*. Students who are neither sick

[4]GK&K and K&S introduce the image concern as either the probability to have told the truth, conditional on the report, or as the probability to have lied, conditional on the report. D&D further interact the conditional probability to have lied with the perceived size of the lie. For example, in D&D the agent gets a lower image if they are suspected of reporting a five instead of a one than if they are suspected of reporting a four instead of a three.

[5]Proposition 7 in AN&R, appendix B, provides some general properties of such a model. Their analysis however remains too general to complement the insights derived from the deed-based image model. Indeed, the result that concludes AN&R's meta-study (Finding 10) cannot distinguish between a model that employs a deed-based image concern and a model that uses a character-based image concern as both account for exactly the same empirical facts (*"Only the Reputation for Honesty + LC [deed-based image] and the LC-Reputation [character-based image] models cannot be falsified by our data"* (AN&R, p. 1144)).

nor in an emergency and excuse themselves from the exam are *shirking*. Professing to be sick when one is not constitutes a lie. Students suffer a fixed cost when lying, which is equal to their moral type t. Moral types t are distributed according to some preference distribution among students and have expectation $E(t)$. A student who is shirking or in an emergency will simulate sickness if the benefits from not writing the exam are higher than the cost of lying, that is, if their type is lower than some threshold $\hat{t}$. Since writing the exam is arguably worse when in an emergency, the critical type threshold $\hat{t}_{emergency}$ is larger than $\hat{t}_{shirking}$. A student who claims to be sick after an emergency has an expected type of $E(t|t \leq \hat{t}_{emergency})$, which is larger than the expected type of a student who lies because they are shirking, $E(t|t \leq \hat{t}_{shirking})$. A student who is actually sick has an expected type $E(t)$, because any randomly chosen student is sick with the same probability. The professor does not observe the real reason of a student who claims to be sick. Therefore, upon receiving a message from a student, it is optimal for the professor to form a posterior expectation about the student's expected moral character by weighing all different potential motives with their empirical frequency;

$$\begin{aligned} E(t|\text{message}) =& \text{P}(sick|\text{message})E(t) \\ &+ \text{P}(emergency|\text{message})E(t|t \leq \hat{t}_{emergency}) + \text{P}(shirking|\text{message})E(t|t \leq \hat{t}_{shirking}). \end{aligned}$$

The posterior expectation after receiving a message will always be lower than the prior expectation (that is, $E(t|\text{message}) < E(t)$). This is because the professor cannot distinguish between truthful and untruthful messages – while sickness is not correlated with moral types, the students who send an untruthful message pool with students who send a truthful message, and those who send the untruthful message are of lower expected type. In line with the idea that individuals want to be perceived of high moral character, a student's reputation is equal to the professor's posterior expectation. Now assume that there is a (potentially pandemic-induced) increase in the probability that a student is sick at the exam date. All things equal, such an increase will increase the professor's posterior expectation. This reflects the credibility effect – if more students are actually sick, it is more likely that any student claiming to be so is telling the truth. Alternatively, consider an increase in the probability that any student faces an emergency at the exam date (which might also be pandemic-induced as they have to care for sick family members). Such an increase will also increase the professor's posterior expectation, as, conditional on not being sick, it is less likely that the student is simply shirking. This reflects the quality effect – even though they may still be liars, students with an emergency have a better reason to lie.

In the die roll game, the character-based model predicts an equilibrium that can include partial lying. Recall that agents have a financial incentive to overstate their number. Therefore, if some agents lie to report the highest paying number, this number will on average be reported by worse moral types. Because agents are image concerned, they might then have an incentive to leave some money on the table in exchange for a higher image by reporting the second-highest or an even lower number when they lie. This dynamic generates an equilibrium with characteristics that are similar to the deed-based image models of GK&S and K&S;

agents lie only if they draw a number that is smaller than or equal to some threshold and report a number that is above the threshold. Under an equilibrium refinement that restricts liars to play symmetric strategies, this is the unique outcome of the game.

With the quality effect, the model introduces a signaling motive whose effect on lying has not been studied before. Its relevance depends on the opportunity cost of truth-telling. In the die roll game, individuals who truthfully report a four, for example, have a different opportunity cost of truth-telling than individuals who truthfully report a one. In both cases, individuals could have lied to report a five. By being truthful, individuals who report a one forego a larger marginal gain than individuals who report a four. Hence, while both statements are honest, a report of one sends a higher signal about intrinsic honesty ("quality") than a report of four. In consequence, this effect leads to new predictions about the strategic complementability and substitutability of lies and about the effectiveness of norm interventions, which I explore in the paper. I also ask what further can be learned from assuming that agents not only differ in their moral character but also in the extent to which they care about their image.

The second half of the paper complements the theoretical analysis with the design and results from a lab experiment. The experiment implements a lying game in which some participants' reasons to lie are more justified than those of others (having either a low or high endowment). The first two treatments exogenously vary whether the average liar either has a worse or a better reason to lie. As lying should be less stigmatized, and thus more attractive, when liars have better reasons, comparison of reporting behavior between both treatments gives an indication of the quality effect. A third treatment exogenously reduces the number of participants who truthfully report the highest state in each session. Compared to the remaining treatments, this variation makes reporting the highest state look more suspicious and hence less attractive. It should give an indication of the credibility effect.

Contrary to expectations, the treatment effects I estimate are very close to zero and thus provide no evidence that image concerns motivate lying. In order to better understand the null results, I further investigate the relation between data I collected on beliefs about lying and behavior. Both reduced-form evidence and structural estimates suggest that, while beliefs and behavior are correlated, beliefs do not causally affect behavior. In contrast, behavior in my experiment seems mainly to be driven by intrinsic motives and lying behavior is responsive to financial stakes.

The following section presents the model. At the beginning, parts 2.1 and 2.2 discuss the setup and equilibrium properties. I apply the model to investigate strategic complementarity and substitutabilty of lies in part 2.3 and provide an extension of the model to heterogenous image concerns in 2.4. Thereafter, in 3, I relate the character-based approach developed in this paper to the relevant theoretical literature and discuss experimental evidence on image concerns and lying. Section 4 presents the experimental design and section 5 the results. The paper closes with a discussion in section 6 that relates the empirical findings to the existing experimental literature.

2 Model

2.1 Setup

Game form Consider a game between a continuum of agents and an observer. Each agent draws a state $j \in \{1,...,K\}$, which is randomly determined by nature. The agents can be thought to be participants in an economic experiment who are asked by the experimenter, who is the observer, to roll a die. In this case, the state would be the outcome of a die roll. An alternative interpretation of the setup could see agents as students who, at the day of an exam, are either sick or healthy and either are in an emergency or not. Throughout this section, we will focus on the first interpretation. In line with the die roll analogy, we make the simplifying assumption that the state is distributed uniformly on its domain.

After the draw, agents each make a report $a \in \mathcal{K} = \{1,...,K\}$ to the observer and receive a total payoff consisting of direct and image payoffs, as described below. The observer is a passive player with no action whose payoff we do not further specify.

Direct payoffs Agents know their state j, and make a report a, which earns them a direct payoff $y(a)$, where $y(a) - y(a-1) = \Delta(a, a-1) > 0$. The payoff scheme might reflect the experimenter's choice of rewards for reporting numbers certain numbers of the die. Alternatively, the agent-as-student would always earn the highest payoff by claiming to be sick and excusing themselves from the exam.

Reporting $a \neq j$, agents incur cost t which is heterogenous across agents. This cost arises through a purely intrinsic, moral preference for honesty. That indiviudals are heterogeneous in their preferences for honesty is documented in experiments such as Gibson, Tanner, and Wagner (2013), Gneezy, Rockenbach, and Serra-Garcia (2013), and Kajackaite and Gneezy (2017). Gibson et al. (2013) in particular show that the lying cost distribution function consists of many intermediate types, who begin to lie if the returns to lying are high enough. The intrinsic preference for honesty reflects that agents feel bad from lying. This cost is unknown to the observer, who however knows that it is drawn from a distribution $F(t)$ with full support on $(0,\bar{t}]$ and which is independent of j. $\bar{t}$ is a large number, to be specified in detail below. The density function $f(t)$ is log-concave and $E(t|t > \hat{t})$ is convex for any $\hat{t} \in (0,\bar{t}]$.[6]

I will use "lying cost" and "moral type" interchangeably when discussing t, as this section considers honesty as the only relevant moral dimension. This is due to the setup of the game, which reflects laboratory lying games and elements of verbal communication. In these settings, lying comes at no expense to a third party, which allows us to exclusively focus on honesty.[7] Further morality dimensions, such as altruism, might become relevant and interact with honesty in settings where agents cheat someone else, for example stealing (footnote 12 in section

[6] Log-concavity is a very common assumption in the signaling literature and the mathematical properties of log-concave distributions are well understood (see Bagnoli and Bergstrom, 2005, for an overview). Convexity of $E(t|t > \hat{t})$ is a more special assumption, but it holds for many commonly used probability distributions, such as the normal and the uniform distribution. The assumption implies that the individual likelihood to lie is concave in the reputation of the highest state.

[7] The setup might further reflect tax reporting, where individual contributions are a negligible part of total tax earnings.

2.3 provides further discussion of this point).

Image payoffs In addition to being intrinsically honest, agents also value having a reputation for honesty. There can be instrumental reasons to value such a reputation. An expert might prefer to appear honest so as to build an enduring relationship with an advisee. A student who hopes receive a good letter of support from their professor wants to appear sincere to them. There are also non-instrumental reasons for why an agent might prefer to look honest; many individuals want to appear moral and one indicator of morality is honesty. This type of image concern follows B&T and other approaches in psychological game theory that formalize the idea that individuals want to signal virtues that make them a "good guy" (Battigalli and Dufwenberg, forthcoming): Through their actions, agents tell others something about their intrinsic preferences, and agents want to look as if they have preferences which are valued by an observer. To make an inference, the observer forms a belief about the expected moral type of an agent reporting a, denoted as $\mathcal{R}_a = E(t|a)$. The image payoff equals the reputation weighted by a scalar $\mu > 0$,

$$\mu \mathcal{R}_a,$$

where μ is not too large, so that agents are not disproportionally sensitive to changes in the image payoff.[8]

Utility Direct and image payoffs add up to total payoffs, or utility. An agent of type (j,t) who reports a earns utility

$$u(j,t,a) = y(a) - 1_{a \neq j} t + \mu \mathcal{R}_a.$$

I now make an assumption on the maximum lying cost which is a number $\bar{t} > \Delta(K,1) + \mu E(t)$. The assumption ensures, in line with the empirical evidence provided by AN&R, that there are agents who never lie, regardless of the state they draw. One immediate consequence of the assumption is that the observer always puts a positive probability on any state being reported. This property is helpful when solving for the equilibrium, as described next.

2.2 Equilibrium

The structure of the game makes it a *psychological game* (Geanakoplos, Pearce, and Stacchetti, 1989, Battigalli and Dufwenberg, 2009), as final payoffs of agents depend on the observer's belief about the agents' moral type. Agents' strategies s map their type into a distribution over reports. Denote the probability of an agent of type (j,t) reporting a by $s(a|j,t)$. In the following, an agent is a *liar* if they have a strategy where $s(a|j,t) > 0$ for some $a \neq j$, even if also $s(j|j,t) > 0$. To put it another way, an agent who does not always tell the truth is a liar. Conversely, an agent tells the truth iff $s(j|j,t) = 1$.

The following equilibrium definition invokes the standard conditions of utility maximization and that agents and the observer correctly apply Bayes' rule and have a common prior. This definition follows the previous literature and serves as a useful yardstick to think through

[8] See also Lemma 1 in appendix A. B&T, in section 3, provide an extensive discussion on how equilibria depend on assumptions on μ and $f(t)$ in the class of signaling models studied by them. Many of their insights translate to the setting studied in this paper.

strategic interdependencies. Since the maximum lying cost is high, every state is reported with positive probability in equilibrium. This implies that Bayes' rule can be applied to calculate the equilibrium reputation of every state, obliterating the need for further equilibrium refinements to pin-down beliefs that are off the equilibrium path.

Definition 1. *An equilibrium is defined by strategies $s(a|j,t)$, where*

- $s(a=j|j,t) \geq 0$, $s(a \neq j|j,t) \geq 0$ *and* $\sum_{k=1}^{K} s(a=k|j,t) = 1$ *for all* j *and* t.
- $s(a|j,t) > 0$ *if and only if* $a \in \underset{a \in \mathcal{K}}{arg\,max}\ y(a) - 1_{a \neq j} t + \mu E(t|a)$.
- *The agent and the observer hold the correct equilibrium beliefs*

$$\mathcal{R}_j = \frac{\sum_{l=1}^{K} \int_0^{\bar{t}} s(j|l,t) t f(t)\,\mathrm{d}t}{\sum_{l=1}^{K} \int_0^{\bar{t}} s(j|l,t) f(t)\,\mathrm{d}t} \text{ for } j \in \mathcal{K}.$$

2.2.1 Equilibrium properties

The following part contains general observations about the properties of any equilibrium. They lay the groundwork for the main prediction which follows afterwards.

We start with an incentive constraint. An agent who draws state j will lie if there is a state k such that

$$y(k) - t + \mu \mathcal{R}_k > y(j) + \mu \mathcal{R}_j. \tag{1}$$

Since $y(K) > y(j)$ for $j < K$, there cannot be an equilibrium where all agents tell the truth. In this case the reputational payoff would not depend on the reported state, and there would be an agent of type (j, ε), where $\varepsilon > 0$ is arbitrarily close to zero, who could gain by reporting K. Because lying costs are fixed, agents always can make a report a to gain a gross payoff before lying costs of size

$$a \in \underset{a \in \mathcal{K}}{\arg\max}\ y(a) + \mu \mathcal{R}_a. \tag{2}$$

These considerations imply the following observation.

Observation 1. *In an equilibrium, any state that is reported dishonestly by some agents must give the same payoff gross of lying costs as specified in* (2).

It is useful to define a set

$$\mathcal{Q} = \left\{ j \in \mathcal{K} | j \in \underset{a \in \mathcal{K}}{\arg\max}\ y(a) + \mu \mathcal{R}_a \right\}$$

that collects all states that are reported dishonestly with positive probability in equilibrium.

A second consequence of the above reasoning is that there can also not be an equilibrium where someone lies and reports a state j if some agent lies after drawing j. If someone who draws j lies, this implies that $j \notin \underset{a \in \mathcal{K}}{\arg\max}\ y(a) + \mu \mathcal{R}_a$. Therefore, no agent will lie and report j if $s(k|j,t) > 0$ for some $k \neq j$. By the same reasoning, no agent will lie if they draw a state $j \in \mathcal{Q}$,

as lying is costly and does not lead to higher payoffs.
The following observation summarizes the considerations above.

Observation 2. *If* $s(k|j,t) > 0$ *for some* $k \neq j$ *then,* (i)*,* $s(j|l,t) = 0$ *for* $l \neq j$ *and,* (ii)*,* $s(k|k,t) = 1$*.*

We now investigate the role of the lying cost has in determining an agent's report. Consider again the incentive constraint (1) and note that the payoff from lying strictly decreases in the lying cost. It follows that an agent lies if their lying cost is sufficiently low. In particular, for each state j there will be a threshold lying cost $\hat{t}_j$ and agents (j,t) will lie if $t \leq \hat{t}_j$, where $\hat{t}_j > 0$ if $j \notin \mathcal{Q}$ and $\hat{t}_j = 0$ otherwise. Now consider the reputations that are associated with agents who draw state j. Truth-tellers comprise the upper tail of the preference distribution, while liars make up the lower tail. Truth-tellers and liars have an expected cost of respectively

$$\mathcal{M}^+(\hat{t}_j) \equiv E(t|t > \hat{t}_j),$$
$$\mathcal{M}^-(\hat{t}_j) \equiv E(t|t \leq \hat{t}_j).$$

The first term is naturally larger than the second, which reflects that liars are stigmatized while truth-tellers are honored. We collect all cost thresholds $\hat{t}_j$ of each state in a vector $\hat{\mathbf{t}}$ and define the *expected lying cost of liars* by

$$\mathcal{L}(\hat{\mathbf{t}}) \equiv \sum_{j \notin \mathcal{Q}} \mathrm{P}(\text{draw } j|\text{lie}) \mathcal{M}^-(\hat{t}_j), \text{ with } \mathrm{P}(\text{draw } j|\text{lie}) = \frac{F(\hat{t}_j)}{\sum_{k \notin \mathcal{Q}} F(\hat{t}_k)}. \tag{3}$$

Turning to the reputation of reporting a state, observation 2 implies that, if a state is not lied at, its reputation is equal to the expected type of agents who are above the threshold;

$$\mathcal{R}_j = \mathcal{M}^+(\hat{t}_j) \text{ if } j \notin \mathcal{Q}. \tag{4}$$

Every state in $\mathcal{Q}$ is reported honestly by a fraction $1/K$ of all agents. In addition, it is also reported by liars. Define by

$$r_j \equiv \mathrm{P}(\text{truth}|\text{report } j) = \frac{1}{1 + \sum_{k \notin \mathcal{Q}} \int_0^{\hat{t}_k} s(j|k,t) f(t)\, \mathrm{d}t}$$

the probability that a randomly chosen agent reporting j is telling the truth. The reputation of reporting j then becomes

$$\mathcal{R}_j = r_j E(t) + (1 - r_j) \frac{\sum_{k \notin \mathcal{Q}} \int_0^{\hat{t}_k} s(j|k,t) t f(t)\, \mathrm{d}t}{\sum_{k \notin \mathcal{Q}} \int_0^{\hat{t}_k} s(j|k,t) f(t)\, \mathrm{d}t} \text{ if } j \in \mathcal{Q}.$$

This expression reflects the reputational spillovers of liars. Since the observer cannot be sure whether any agent reporting j is a liar or not, her best guess is to average the types of those agents who report j honestly and those who lie, weighted by r_j. Pooling with liars spoils the image of truth-tellers, as they are suspected to be the kind of agent who lies. For liars, pooling is image-enhancing, as they cannot perfectly be detected.

One consequence of the results above is that some agents always lie to report K, which gives

the highest direct payoff. Intuitively, every state that is only reported by truth-tellers gives an image payoff above the observer's prior belief (that is, $\mathcal{R}_j > E(t)$ if $j \notin Q$). On the other hand, states that are reported by liars reduce the observer's prior *on average*. Therefore, there cannot be an equilibrium where K gives the highest direct payoff and increases the observer's prior belief because such a situation makes it too attractive for liars to report K and experience an increase in the direct and the image payoff. By similar arguments, no agent will ever lie to report 1, the state which pays the lowest direct payoff:

Observation 3. *In equilibrium,* (i) $K \in Q$ *and* (ii) $1 \notin Q$.

Proof. (i) Assume the contrary, $K \notin Q$. Then, for all states $j \in Q$,

$$y(j) + \mu\mathcal{R}_j > y(K) + \mu\mathcal{R}_K, \text{ and } y(K) > y(j). \tag{5}$$

This in particular implies that $\mathcal{R}_j > \mathcal{R}_K$ for all $j \in Q$. From (4) it follows that $\mathcal{R}_K > E(t)$ and more generally $E(t|\text{report } j \notin Q) > E(t)$. By the martingale property of beliefs, it then follows that $E(t|\text{report } j \in Q) < E(t)$, which requires that $\mathcal{R}_j < E(t)$ for some $j \in Q$.[9] Combining the inequalities, we arrrive at $\mathcal{R}_K > E(t) > \mathcal{R}_j$ for some $j \in Q$, which is a contradiction to (5).

(ii) Assume the contrary, $1 \in Q$. Then, for all states $j \notin Q$,

$$y(j) + \mu\mathcal{R}_j < y(1) + \mu\mathcal{R}_1, \text{ and } y(1) < y(j). \tag{6}$$

This in particular implies that $\mathcal{R}_1 > \mathcal{R}_j$ for all $j \in Q$. Since $\mathcal{R}_1$ is a convex combination of the prior and the reputation of liars, the highest value $\mathcal{R}_1$ can obtain is smaller than $\max\{E(t), \max\{\hat{\mathbf{t}}\}\} < E(t|t > \max\{\hat{\mathbf{t}}\})$. Since $\mathcal{R}_j = E(t|t > \max\{\hat{\mathbf{t}}\})$ for some $j \in Q$, we arrive at a contradiction to (6). ■

Some agents are always dishonestly reporting K, but is this also the only state that will be reported dishonestly in equilibrium? The answer to this question is no, if the direct payoff becomes small relative to the image payoff.

Observation 4. *Consider the case where* $K > 2$. *If the ratio* $\frac{\Delta(K,K-1)}{\mu}$ *is sufficiently small, then there is no equilibrium where* $Q = \{K\}$.

Proof. Consider an equilibrium where Q is a singleton. It then holds that

$$y(K-1) + \mu\mathcal{R}_{K-1} < y(K) + \mu\mathcal{R}_K,$$

because every liar must prefer to report K over $K-1$. We can rearrange this inequality to

$$\mathcal{R}_{K-1} - \mathcal{R}_K \leq \frac{\Delta(K,K-1)}{\mu}. \tag{7}$$

[9] The martingale property states that a Bayesian observer never changes her prior on average. In the present context, $E[E(t|a)] = E(t)$.

Since $K-1 \notin \mathcal{Q}$, it follows from (4) that $\mathcal{R}_{K-1} > E(t)$. Furthermore, if $\mathcal{Q}$ is a singleton then

$$\mathcal{R}_K = \frac{1}{1+\sum_{j\notin\mathcal{Q}} F(\hat{t}_j)} E(t) + \frac{\sum_{j\notin\mathcal{Q}} F(\hat{t}_j)}{1+\sum_{j\notin\mathcal{Q}} F(\hat{t}_j)} \mathcal{L}(\hat{\mathbf{t}}) < E(t).$$

The left-hand side of (7) is strictly positive. Thus, there is a contradiction if $\frac{\Delta(K,K-1)}{\mu}$ is sufficiently small. ■

The intuition behind this result is that, with image concerns, liars trade off a higher direct payoff with a lower image payoff. Since liars spoil the reputation of the states they report, it is beneficial for them to spread out and report more than one state to "smooth out" the image loss over multiple states.

2.2.2 Equilibrium refinement, main prediction

The predictions above can be useful, but they are also relatively unspecific. One reason is that the equilibrium definition allows for a very rich variety of strategies that liars can play, some of which that might appear "strange", or, at least, would require a considerable amount of coordination among liars. For example, with $K=4$, there can be an equilibrium with $\mathcal{Q}=\{2,4\}$, where some liars from 1 lie up to report 2 and some agents from state 3 lie down and to also report 2. This equilibrium can be sustained if liars coordinate on quality; that is, the liars with the highest intrinsic cost report 2 while those with the lowest intrinsic cost report 4. Such behavior can be seen as problematic. In equilibrium, liars are indifferent between reporting any of the states in $\mathcal{Q}$ and there is no a-priori reason why some liars would prefer to report one state over another. The degree to which liars have to coordinate to support such an equilibrium motivates a refinement that restricts agents to symmetric lying strategies, as defined below.[10]

Definition 2. *Agents play symmetric lying strategies if* $s(k|j,t), s(k|j',t') > 0 \Rightarrow s(k|j,t) = s(k|j',t')$ *for any* $t,t' \in (0,\bar{t}],\ j,j' \notin \mathcal{Q}$.

Put differently, lying strategies are symmetric when the agents' type (j,t) determines whether they report a state in $\mathcal{Q}$ or not, but does not determine which state in $\mathcal{Q}$ they report. A similar property ("uniform cheating") is imposed by D&D to obtain their main result. Symmetrict lying strategies imply that liars randomize which state to report dishonestly. While there are few direct tests of mixed lying strategies, evidence from F&FH is seemingy in line with this refinement. They show that participants who participate in a die-roll experiment for a second time, and who reported the second-highest or highest state in the first experiment, show exactly the same average behavior in the second experiment. If liars had further conditioned their reports on some intrinsic attributes, we would expect reports of those who report the highest state to be systematically different from those who report the second-highest state.[11]

Solving the model under symmetric strategies gives the main result.

[10] Appendix B gives a numerical example of an asymmetric equilibrium.

[11] F&FH also show that participants who report lower numbers in the first experiment are more likely than others to report lower numbers in the second experiment, implying that decisions are to some extent consistent across both experiments.

Proposition 1. *There exists a unique equilibrium when agents play symmetric lying strategies. The equilibrium has the following properties:*

(i) *The report distribution is strictly increasing in* j.

(ii) $\mathcal{R}_j$ *is strictly decreasing in* j.

(iii) *No agent who draws* j *reports a state* $k < j$.

(iv) $Q = \{j \in \mathcal{K} | j > k^*\}$, *where* $k^* \in \mathcal{K} \setminus \{K\}$.

The equilibrium of the game is of the following type: Agents lie only if they draw a state smaller or equal than some threshold state k^*. If they lie, they report a state larger than k^*. State K is reported by most agents, followed by $K-1$, and so on. In what follows, I provide a sketch of the proof and relegate the technical details to appendix A.

With symmetric strategies, the reputation of states $j \in Q$ becomes a weighted average between the quality of non-liars and the expected lying cost of liars;

$$\mathcal{R}_j = r_j E(t) + (1 - r_j)\mathcal{L}(\hat{\mathbf{t}}) \text{ if } j \in Q. \tag{8}$$

One immediate consequence of this is that there is no downwards lying (part (iii) of proposition 1). To see that, note that we now have $\mathcal{R}_j < E(t) < \mathcal{R}_k$ for all $j \in Q$, $k \notin Q$. Reporting a lower state than the one one has drawn now would imply both, a lower image and a lower direct payoff, which is inconsistent with utility maximization. Part (iv) of the proposition is a direct implication of part (iii).

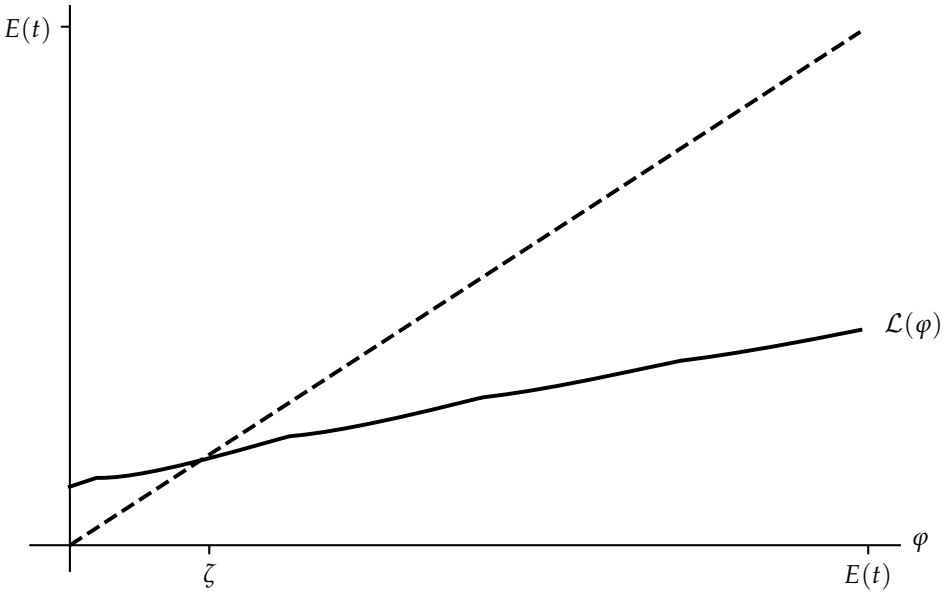
Turning to reputations, it is useful to distinguish between states that are not in Q and those that are. Among the former, reputations decrease the higher the state because agents in higher states have lower direct incentives to lie. For example, agents who report 1, despite having a high incentive to lie, send a higher signal about their intrinsic honesty than agents who report k^*. Reputations also intuitively decrease among states in Q, because liars trade off direct payoffs for image payoffs. The lower the direct payoff of a state, the more the liar has to be compensated with a higher image payoff.

Decreasing reputations imply increasing reporting frequencies; among those states that are not being lied at, there is an inverse relation between the reputation of the state and the proportion of agents who report it. With symmetric lying strategies, the same relation holds among states that are being lied at, as the reputation of any state is decreasing in the proportion of liars that are reporting it. Therefore, in the proposition, (i) is a consequence of (ii).

Constructing the equilibrium is seemingly complicated, because it involves a threshold state k^*, a vector of threshold costs $\hat{\mathbf{t}}$, and a vector of probabilities $\mathbf{r} = (r_{k^*+1}, ..., r_K)$ that each depend on one another. The key step in the proof is to realize that we can fix the reputation of the highest state, which is always reported dishonestly in equilibrium, at some level φ. We can then define a function which implicitly defines the cutoff lying cost *as a function of* φ;

$$\mathcal{T}(\hat{t}, \varphi, \Delta(K, j)) \equiv \Delta(K, j) + \mu[\varphi - E(t | t \geq \hat{t})] - \hat{t} = 0. \tag{9}$$

Figure 1. Lying cost of liars



Note: The dashed line represents the 45 degree line.

Since $E(t|t \geq \hat{t})$ is monotonly increasing in $\hat{t}$, there is always a unique cutoff for a given φ, which we denote as $\hat{t}(\Delta(K,j),\varphi)$ (or sometimes also as $\hat{t}_j(\varphi)$ to save notation).
Plugging these functions into (3), the expected lying cost of liars becomes a function of φ only;

$$\mathcal{L}(\varphi) = \sum_{j \leq k^*} \frac{F(\hat{t}_j(\varphi))}{\sum_{k \leq k^*} F(\hat{t}_k(\varphi))} \mathcal{M}^-(\hat{t}_j(\varphi)).$$

We can then observe that the equilibrium reputation of the highest state, φ^*, must be between $\mathcal{L}(\varphi^*)$ and $E(t)$, as the reputation is a weighted average of the expected lying cost of liars and that of truth-tellers. Possible values of φ that are consistent with equilibrium are on the domain between the *fixed-point* where $\xi = \mathcal{L}(\xi)$ and $E(t)$, as displayed in figure 1.

The functions $\hat{t}_j(\varphi)$ can be thought of as characterizing the supply of lies, as $F(\hat{t}_j(\varphi))$ gives the proportion of agents from state j that are lying. Aggregating them up in a function

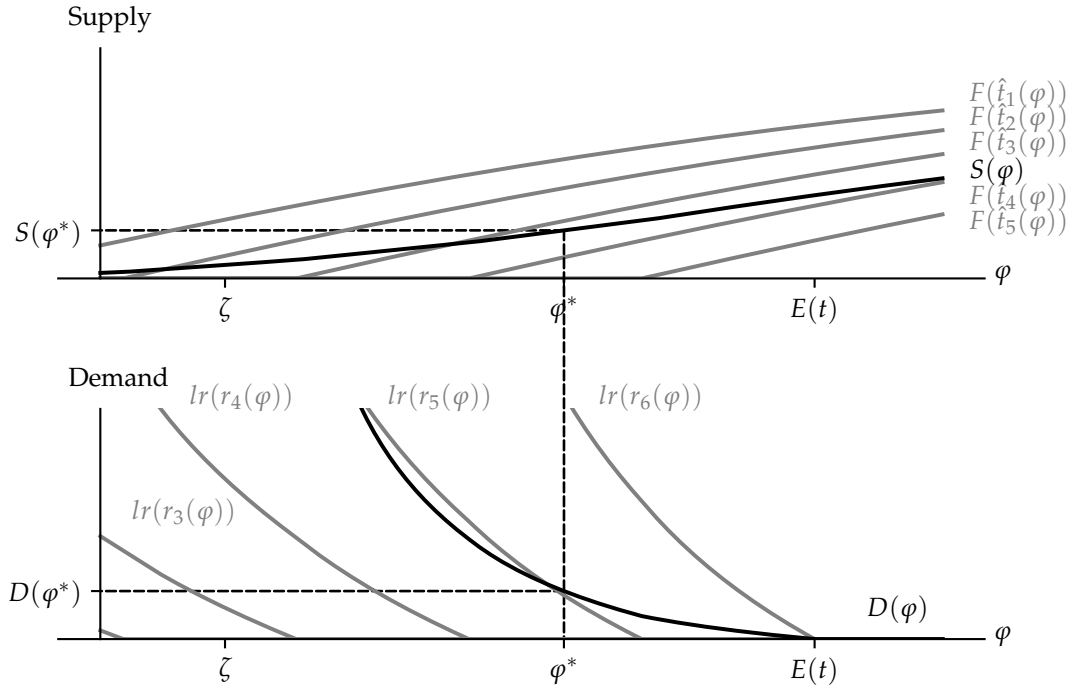
$$S(\varphi) \equiv \frac{1}{K} \sum_{j=1}^{k^*} F(\hat{t}_j(\varphi))$$

gives the fraction of agents that are willing to lie if the reputation of the highest state is φ.

We can similarly rearrange (8) to write r_K as a function of φ. The indifference condition from observation 1 allows us to derive a function $r(\Delta(K,j),\varphi)$ for all remaining states $j \in \mathcal{Q}$.

The $r_j(\varphi)$ functions characterize the demand side. Transforming $r_j(\varphi)$ to a likelihood ratio $lr_j(\varphi) = \frac{1-r_j(\varphi)}{r_j(\varphi)}$ gives the ratio of liars to non-liars reporting j if the reputation of the highest

Figure 2. Equilibrium



state is φ. Adding up the likelihood ratios and normalizing by $1/K$, we arrive at a function

$$D(\varphi) \equiv \frac{1}{K} \sum_{j=k^*+1}^{K} lr_j(\varphi).$$

This function returns the proportion of agents who lie, as a function of φ. It can be interpreted as a demand function, as it gives the fraction of liars that are needed to sustain an equilibrium for a given reputation of the highest state.

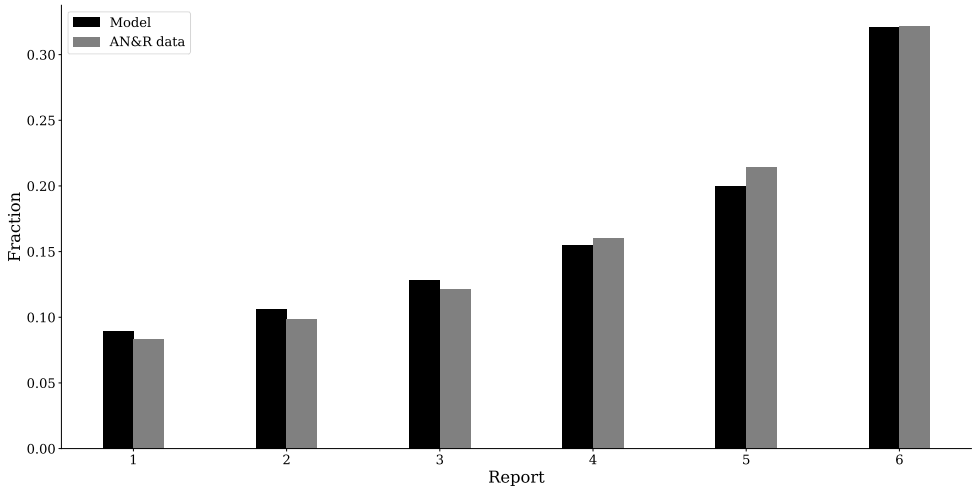
Figure 2 illustrates both functions. The upper panel shows the individual threshold functions and the aggregate supply function. An increase in the reputation of the highest state makes it more attractive for agents to lie, which is why the function slopes upwards. The lower panel shows the demand side. These functions slope downwards. Intuitively, when φ is low, many liars will report states different from K to alleviate reputational losses. But such behavior requires that a high proportion of agents lies to sustain the indifference conditions. Conversely, as φ approaches $E(t)$ every liar will report the highest state, which is possible only if a small proportion of agents lies.

Supply and demand have to coincide in equilibrium, which determines φ^*, which in turn pins down $\hat{\mathbf{t}}$, $\mathbf{r}$, and k^*. The regularity conditions imposed on $F(t)$ ensure existence and uniqueness of φ^*.

Equilibrium behavior is shaped by signaling motives and a number of insights follow:

Reporting in equilibrium The equilibrium predicts a report density that is increasing in the direct payoff. We would obtain the same prediction from a model with only intrinsic lying costs. The key difference between both models is however that, when they are image concerned,

Figure 3. Example equilibrium report distribution compared to the AN&R data



Note: Example equilibrium distribution of reports when lying costs follow a log-normal distribution where log-costs have mean zero and standard deviation 1.1, for values $y(a) - y(a-1) = 1$, and $\mu = 2.1$.

some agents might lie and report a non-payoff maximizing state. For example, the model can match the empirical findings from die roll games in laboratory experiments quite well. Figure 3 compares the predicted equilibrium distribution for a calibrated version of the model to the data collected by AN&R. The model comes close to the observed frequency distribution and in particular can account for partial lying.

Freeriding on reputation Liars report a state different from the highest state only if they get a higher image payoff in return. One necessary condition for this image enhancing effect is that every state which does not maximize direct payoffs, and which is reported by a liar, is also reported honestly by some agents. In equilibrium honest agents and liars pool, and liars free-ride on the honest agents' reputation.

Image spoiling mechanisms States that are reported dishonestly suffer a reputational penalty because of two factors; credibility and quality. If a state is reported by many liars, then any single agent reporting that state does not appear to have credibly done so truthfully. Since liars are of a worse moral type then the average agent, the reputation of a state suffers when more liars are reporting it. In addition, the reputation of a reported state also depends on the kind of liars that are reporting it. Liars are of higher reputation if they have a relatively high lying cost. That is, they only lie if there are substantial utility gains from lying that give them good reasons to lie. The marginal liar in state j, who is of type $\hat{t}_j$, always has a higher reputation than the inframarginal liars, who are of expected type $\mathcal{M}^-(\hat{t}_j)$. This implies that the quality of liars is increasing in their amount. In the limit as $\hat{t}_j \to \bar{t}$, then $\mathcal{M}^-(\hat{t}_j) \to E(t)$; there is no stigma associated with liars from state j. This reflects that bad behavior can be normalized because "everybody is doing it". If almost all agents are committing the bad deed, then doing so oneself is no longer a sign of low character, but merely a signal of mediocrity.

2.3 Actions: strategic substitutes or complements?

Many information nudges that address unethical behavior such as tax evasion rely on norm reminders. The main idea is that such reminders, for example about aggregate tax compliance, might be useful in convincing individuals to also comply with the majority of individuals who are compliant. Such interventions implicitly assume that individual actions are strategic complements. This part investigates the theoretical determinants of strategic complementarity and substitutability of lies.

In the model, the credibility of a report decreases in the number of liars while the quality of the report increases in the number of liars. There is a tension between both effects. This tension can lead to tradeoffs that affect whether lies are strategic complements or substitutes. We will examine these strategic interactions by exogenously shifting the type of the marginal liar in state j and evaluating its effect on lying from other states. Other states are affected by changes in $\hat{t}_j$ because such changes have an impact on the image payoff from lying. If the reputation of the highest state increases in response to an increase in $\hat{t}_j$ then agents from other states will be encouraged to lie. Otherwise, they will be discouraged. Formally, we have the following definition.

Definition 3. *Lies are strategic complements with respect to the jth state if* $\frac{\mathrm{d}\varphi^*}{\mathrm{d}\hat{t}_j} > 0$.

Define $S(\varphi,\hat{t}_j)$ and $D(\varphi,\hat{t}_j)$ as in proposition 1 but where the lying threshold of one state $\hat{t}_j$ is determined exogeneously. Consider an increase in lying from agents who draw state j, which results in an increased lying cost threshold from $\hat{t}'_j$ to $\hat{t}''_j > \hat{t}'_j$. After the increase, the supply of lies becomes larger – at any given level φ more agents are lying ($S(\varphi,\hat{t}''_j) > S(\varphi,\hat{t}'_j)$). This reflects the mechanical credibility effect, which makes it more likely that any agent reporting a high state is a liar. If isolated, this effect crowds out lying of agents from all other states. In addition, there is a quality effect whose sign is not unambiguously determined. If the average liar is of better quality after the increase, lying becomes more attractive. Everyone then benefits from the higher reputation that is associated with lying. In this case, the quality effect goes against the credibility effect. There can however also be situations where the quality and credibility effects go into the same direction, namely if an increase in $\hat{t}_j$ makes the average liar worse. In general, an increase in $\hat{t}_1$ will always lead to a positive quality effect, while an increase in $\hat{t}_{k^*}$ makes it more likely that the quality and credibility effect are both negative. Intuitively, liars who draw state 1 have the best, and liars who draw k^* have the worst reasons to lie. Liars in k^* are relatively cheap; they lie for a smaller utility gain than liars in 1, which makes them the lowest quality liars on average.[12]

Bénabou and Tirole (2011) show that norm reminders can be effective if agents have image concerns and are uncertain about the distribution of preferences in the population. Through

[12]The fact that "small" lies are more severely stigmatized than "large" lies would be more ambiguous in a setting where agents' lying decisions have direct payoff implications for a third party. In settings where agents cheat at the expense of others, it would be appropriate to introduce further moral dimensions, such as pro-sociality, into the model. The consequence might be that a "large" lie is more stigmatized than a "small" lie, because agents who take from someone else signal that they care little about the welfare of others. (See e.g. Cohn, Maréchal, Tannenbaum, and Zünd (2019) for further discussion and evidence that individuals are less likely to cheat for a large gain than for a small gain when they believe that someone else will suffer from it.)

revealing that agents on average are sufficiently moral, a planner can increase aggregate moral behavior in the population. A number of experiments suggest that participants react in conformist ways to information about past report distributions (Rauhut, 2013, Diekmann, Przepiorka, and Rauhut, 2015), indicating that lies are strategic complements. These effects are generally small and have not been confirmed by AN&R, who in one experiment find an insignificant effect of shifting beliefs about lying on behavior. From a theoretical perspective, one interesting aspect in lying games is that the bad deed (lying) cannot be perfectly identified. This can hamper the effectiveness of norm reminders and might explain its ambiguous empirical effects. Through the lens of character-based image models, norm reminders work by informing agents that on average, the population is sufficiently honest. This increases the stigma from lying and therefore encourages honesty through the quality effect. However, information that only few agents lie also indicates that reporting the high state will not raise many suspicions. Therefore, norm reminders might work worse to discourage behavior that is only imperfectly observed, such as lying.[13]

2.4 Heterogenous image concerns

This section considers signaling behavior under heterogenous image concerns. We introduce heterogenous image concerns into the model by assuming that agents can be of two image-types (unknown to the observer), a high type (h-type) and a low type (l-type). High types care more about their image than low types. More formally, $\mu_\theta \in \{\mu_l, \mu_h\}$ with $\mu_h > \mu_l \geq 0$. The proportion of l-types in the population is $\rho < 1$. The lying cost of type θ is distributed according to $F_\theta(t)$, which is common knowledge. If $F_l(t) = F_h(t)$ preferences are uncorrelated with the image concern. They can also be correlated. If, for example, $F_l(t) > F_h(t)$ for any t, then agents who are more image concerned also have higher lying costs. We denote the set of states that are reported dishonestly by θ-agents as Q_θ.

The following proposition summarizes features of an equilibrium where agents have heterogenous image concerns.

Proposition 2. *With heterogenous image concerns, $K \in Q_l$. The intersection between Q_l and Q_h is either empty or a singleton. When the intersection $Q_l \cap Q_h$ is a singleton, $Q_l \cap Q_h = min(Q_h)$.*

Agents with high image concerns always report (weakly) lower states when they lie than agents with low image concerns. This happens precisely because lower states pay higher image payoffs.

To illustrate the new features that can arise in an equilibrium with heterogenous image concerns, we will consider two examples of a simple die roll game with only three different payoff levels, where reporting any number lower than five pays one, and reporting five or six pays five and six, respectively. We also keep heterogeneity in the image concern as simple as

[13] Indeed it can be shown that, in the current setting with added preference uncertainty, a norm reminder as in Bénabou and Tirole (2011) is always less effective in promoting honesty in a lying game compared to an observed lying game where the observer knows both, draws and reports.

possible, by assuming that l-type agents are *homines oeconomici* with $\mu_l = 0$, who do not incur a moral cost from lying.[14]

Example 1. (Separaration by image type) *Consider the simple die roll game as described above with $\mu_h = 2$ and $t \sim U[0,8]$. If a fraction $\rho = .1$ are homines oeconomici, there is an equilibrium with no downwards lying and where all h-type agents who lie report five.*

Proof. We construct the equilibrium above. In any equilibrium l-type agents always report six, which pays the highest direct payoff (in the examples we slightly abuse notation by naming the possible states 1, 5, and 6.). The claim in the example then implies that, because there is no downwards lying, that

$$\mathcal{R}_1(\hat{t}) = \frac{\hat{t}+\bar{t}}{2},\ \mathcal{R}_5(\hat{t}) = \frac{1}{2} \times \frac{4\hat{t}^2+\bar{t}^2}{4\hat{t}+\bar{t}},\ \mathcal{R}_6 = \frac{(1-\rho)\bar{t}/2}{1+5\rho},$$

where $\hat{t}$ is the lying cost threshold function, which is endogeneously determined through

$$\hat{t} = \Delta(5,1) + \mu(\mathcal{R}_5(\hat{t}) - \mathcal{R}_1(\hat{t})).$$

Under the uniform distribution assumption, this is a quadratic equation which is solved by

$$\hat{t} = \frac{-\bar{t}(1+\frac{5}{2}\mu) + 4\Delta(5,1) + \sqrt{(\bar{t}(1+\frac{5}{2}\mu) - 4\Delta(5,1))^2 + 16\Delta(5,1)\bar{t}}}{8}, \tag{10}$$

or, for the parameters above, $\hat{t} \approx .9$. Finally, we have to verify that h-type agents are indifferent between reporting 5 and 6, that is if

$$y(5) + \mu\mathcal{R}_5 = y(6) + \mu\mathcal{R}_6. \tag{11}$$

This holds with equality when $\rho = .1$. ■

There are two takeaways from this example. First, it is possible that liars separate by image type, that is, agents with the lowest image concern lie to report the highest state and other agents only lie partially. This is close to a type-based explanation that was suggested by FFH, where some individuals are full liars, others partial liars, and the rest are truth-tellers. However, as the example shows the classification into different types of liars is endogeneously determined by the image concern.

A second insight is that any increase in μ, e.g. by making the lying game more observed, in the setup above would lead to downwards lying. To see this, note that in the equilibrium above, it is reputationally more attractive to report five than six since the homines oeconomici spoil the reputation of the highest state. Therefore, any increase in μ would lead to strict inequality in (11); h-type agents would strictly prefer to report 5 over 6.

Example 2. (Non-modality of the highest state) *Consider the simple die roll game as described above where h-types have $\mu_h = 3/2$ and draw their lying cost from a discrete distribution $(p$:*

[14] This approach follows Grossman and van der Weele (2017), who study the impact of introducing the homo oeconomicus on prosocial behavior in a character-based image model.

$\underline{t}; (1-p) : \bar{t})$, *where* $p = .5$ *and* $\bar{t} = 3.5 > \underline{t} = 3$. *If* $\rho = .21$, *there is an equilibrium with no downwards lying and where more agents report five than six.*

Proof. The equilibrium described above is consistent with the following reputations;

$$\mathcal{R}_1 = \bar{t},\ \mathcal{R}_5 = \frac{5p\underline{t} + (1-p)\bar{t}}{1+4p},\ \mathcal{R}_6 = \frac{(1-\rho)(p\underline{t} + (1-p)\bar{t})}{1+5\rho},$$

which imply the incentive constraints

$y(5) + \mu\mathcal{R}_5 - \underline{t} > y(1) + \mu\mathcal{R}_1$ (h-type agents with low lying costs prefer lying to 5 over honestly reporting 1),
$y(5) + \mu\mathcal{R}_5 - \bar{t} < y(1) + \mu\mathcal{R}_1$ (h-type agents with high lying costs prefer honestly reporting 1 over lying to 5),
$y(6) + \mu\mathcal{R}_5 = y(6) + \mu\mathcal{R}_6$ (h-type agents are indifferent between reporting 5 or 6).

By plugging in the parameter values, it can be verified that they indeed hold. In equilibrium, the fraction of agents reporting six is

$$(1-\rho)/6 + \rho \approx .34$$

and the fraction of agents reporting five is

$$(1-\rho)/6 + (1-\rho)p \times (4/6) \approx .40.$$

Therefore, more agents report five than six. ■

This example shows that, with heterogenous image concerns, the partial lying motive can lead more agents to report the second highest state over the highest state. This is a consequence of the quality signaling motive. Among the different quality "segments" of liars, the homines oeconomici have the worst reputation. Image concerned agents might then be induced to report a state different from the highest state because they do not want to be mistaken for a homo oeconomicus, even if doing so is more obviously a lie.

There is evidence that the highest state is not always reported by most participants. Out of 24 papers included in the AN&R meta-study that employ a one-shot die-roll lying game, 8 papers contain experiments where the highest state is not the modal report. Most of these experiments have been conducted outside of traditional lab environments in settings where the social distance between observer and participants is arguably lower and where the social image motive thus might play a greater role. For example, Ruffle and Tobol (2017) conduct an experiment with Israeli soldiers who have to report the outcome of a die roll to an army official. The higher the reported die roll, the earlier the soldiers will be released from duty at one weekday afternoon. They find that some soldiers lie to the army official and that most of them report the second-highest state.

3 Related literature

This section provides a discussion of the relation between the model and previous theoretical work. Thereafter, I consider experimental research on image concerns and lying.

3.1 Relation to other image models

Agents in the model presented in this paper are motivated by ethics of virtue, while agents in the deed-based models of GK&S and K&S are deontological in the sense that their ethic follows a rule ("you should not lie"). While signaling motives in the deed-based models are only about *credibility* (how suspicious the report is), the character-based model adds a further *quality* effect (what kind of agents are going to lie).

One behavioral difference that arises is that in deed-based image models, lies are strategic substitutes: if the fraction of liars increases, any single lie becomes more suspicious, which discourages individuals from lying through the credibility effect. In the character-based image model, lies can become strategic complements if the quality effect dominates the credibility effect.

With character-based signaling, the goal of the observer is to find out the agent's preferences. In deed-based models, individual preferences do not interact with the signaling motive. In the model proposed by AN&R, for example, all agents have the same, commonly known, intrinsic lying cost. In such an environment the character-based signaling motive loses its bite. The observed lying game provides another illustrative special case. In this game, the observer knows the draw and the report of all agents. With the deed-based approach, lying decisions of any single agent are not influenced by the strategies of others. With the character-based approach, agents' strategies are still interdependent in observed lying games because preferences remain unobserved.[15]

AN&R provide a discussion of deed-based signaling under heterogeneous image concerns. As in the character-based model, downwards lying can arise as part of the equilibrium. As a consequence, an increase in fraction of agents reporting the highest state does not necessarily imply that more agents are lying. Instead, the increase might be driven by less agents lying downwards. An agent who initially underestimates the fraction of agents reporting the highest state might therefore become more likely to lie after learning the true fraction, *despite* lies being strategic substitutes in the model. Intuitively, the agent believes that they initially overestimated the amount of downwards lying, which leads the agent to increase their credibility belief, which in turn makes reporting the highest state more attractive. As this discussion highlights, an empirical test which aims to investigate the strategic complementarity and substitutability of *lies* by providing individuals with information about *reporting frequencies* might only provide ambiguous information, because reporting frequencies might not map one-to-one into lies.

[15] As the deed-based approach, the character-based approach would predict that all liars pool on the highest state in a symmetric equilibrium. This is because partial lies can only be sustained if lying partially provides a higher image. This is impossible in the observed game with symmetric strategies.

D&D consider a setup with no intrinsic lying costs but where image depends on some perceived, increasing lying cost that the observer assigns to judge the intensity of unethical behavior. Thus, observers distinguish between more than simply a good deed (truth-telling) and a bad deed (lying), by interacting actions with the intensive margin of the lie. This could for example be a realistic assumption if observers dislike the deceptive element of a lie, as a large lie (reporting a five instead of a one) intends to deceive the observer by more than a small lie (reporting a four instead of a three). As in the character-based model, incorporating an intensive margin leads to reputations that are strictly smaller the higher the report. The observer's motivation is however different; in D&D, the primary motivation of the observer is to distinguish between acts of lying that might be more or less deceptive, ignoring what the act tells her about the agent's character.

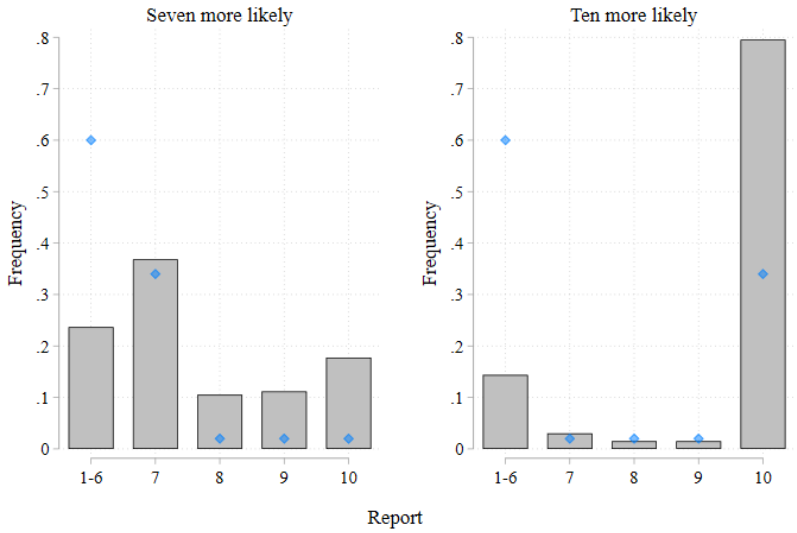
3.2 Experimental evidence

A number of experimental tests show that the belief-dependent models provide correct predictions about lying behavior in previously untested environments. In one of their experimental treatments, GK&S reduce the probability with which participants draw, and therefore truthfully can report, the highest state. The theoretical prediction is that more participants will lie to report a non payoff-maximizing state because it is less credible that participants truthfully report the highest state. The experimental results are in line with this prediction. In similar spirit, AN&R find that participants who draw the lower state in a two-state lying game become less likely to lie when the probability of drawing the high state decreases. Feess and Kerzenmacher (2018) test a related mechanism. In their experiment, they exogenously vary the probability with which participants who toss the lower-paying side of a virtual coin can lie and report the higher-paying side. That is, some participants who toss low can lie while others can not. They find that a smaller proportion of participants lies if there are more participants who have the possibility to lie. This is also consistent with the notion that individuals care about how credible their report is.

In the online appendix to their paper, AN&R present data from an intriguing second experiment on image concerns and lying. In this experiment, participants randomly draw one out of ten numbers and receive a higher payoff the higher the number they report. The distribution of the ten states is non-uniform and over two treatments, the authors vary whether the most likely state is either a seven or a ten, keeping the probability of drawing any of the other states constant. Figure 4 presents the report distribution observed in the experiment. The results indicate that participants who draw a number lower than seven become more likely to lie when probability mass is shifted from seven to ten. Furthermore, participants also tend to tell larger lies – partial lying is absent from the treatment where drawing a ten is more likely than drawing a seven.

Interestingly, the credibility and quality channels discussed in this paper might both contribute to the treatment difference observed by AN&R. The credibility channel will likely matter because, when the likelihood of drawing ten increases, reporting ten appears more credible

Figure 4. Experimental data from AN&R's F10_LOW and F10_HIGH treatments



Note: The histogram shows the distribution of reports for two (between-subject) treatments conducted by AN&R, with a total of 284 participants. The blue dots illustrate the expected distribution of draws. The data used in this figure is available under https://doi.org/10.3982/ECTA14673.

and participants have smaller incentives to disguise their lie by lying partially. In addition, the increase in the probability of drawing a ten comes at the expense of a decrease in the probability of drawing a seven. Therefore, the composition of liars could also be affected by the treatment; participants might believe that liars are more likely to have drawn a number smaller than seven in the treatment where ten is more likely than in the treatment where seven is more likely. This in turn would increase the average quality of liars and thus the perceived attractiveness of lying. Both channels predict the observed treatment effect and could thus jointly account for the observed data.

4 Experiment

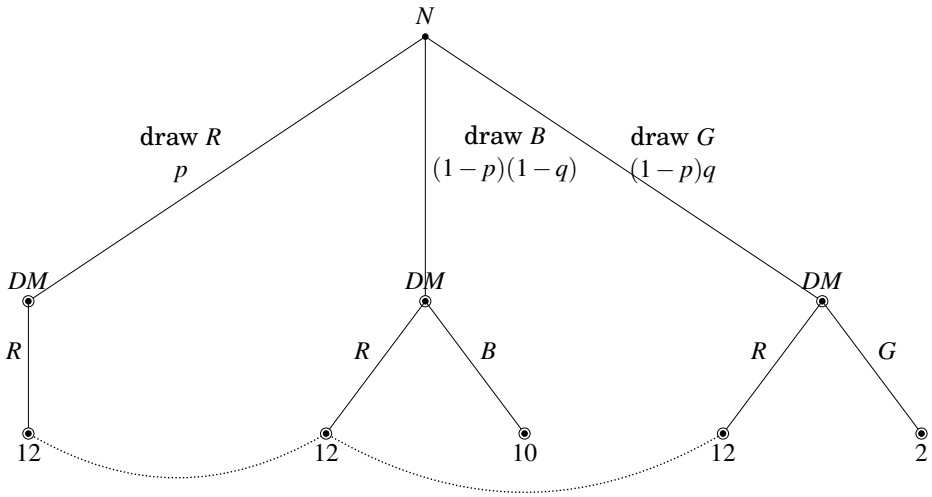
I design an experiment to test the main implications of the model. The reasons for doing an experiment are twofold. First, while there are a number of experiments that test for credibility, the quality channel has not been investigated empirically so far. Second, in addition to measuring behavior, the experiment also elicits beliefs about lying. This way, I can establish whether actions are systematically related to beliefs in ways suggested by the image motive.

4.1 Design

4.1.1 Game

The experiment employs a three-state lying game. Throughout this section, and in accordance with the language used in the experiment, the three states are called *Red*, *Blue*, and *Green*. *Red* is the highest state and reporting it results in a monetary payoff of 12 euros. Reporting

Figure 5. Restricted lying game



Blue gives a slightly lower payoff of 10 euros and reporting *Green* results in 2 euros. To gain more control about the choices liars can take, the lying game is restricted. Figure 5 displays the extensive form of the restricted game. The restricted game differs in two respects from the F&FH lying game. First, partial lying is ruled out by restricting the choice set of potential liars to either report the truth, or to lie and report *Red*. This feature gives control about the actions liars can take, and simplifies the analysis as it rules out the intensive lying margin. Second, the states are not uniformly distributed but drawn with probabilities p (*Red*), $(1-p)q$ (*Green*), and $(1-p)(1-q)$ (*Blue*).

In the experiment, the main treatment variation comes from changing the probability of either drawing the high state (a change in p) or one of the lower states (a change in q). These variations address the two different signaling channels. Consider first a change in p. When decreasing p, the proportion of agents that lie for any given reputation of the highest state, φ, is higher, because less agents draw *Red*. At the same time, the composition of liars does not change. That is, there is only a credibility but no quality effect. This will make reporting the high state ultimately less attractive. Because fewer participants draw *Red*, reporting it appears less credible.

Proposition 3. *A decrease in the probability of drawing Red, p, leads to a decrease in* φ^*.

Turning to comparative statics with respect to q, things get a bit more complicated, because a change in q induces both, a change in the amount and in the composition of liars. The following proposition consists of two parts, with part (i) characterizing the general case and part (ii) giving a sufficient condition for which the quality effect dominates the credibility effect.

Proposition 4. *If the probability of drawing Green, q, increases, then*

(i) $\mathrm{P}(lie)$ *increases.*

(ii) $E(t|\hat{t} \in (\hat{t}_B(\varphi^*), \hat{t}_G(\varphi^*))) > E(t)$ *is a sufficient condition for* φ^* *to increase.*

What intuitively matters for the quality effect to dominate the credibility effect is whether the initial amount of lying is high or low. If the initial amount of lying is high enough, the effect of an improved quality of liars dominates the increased amount of liars. In this case reporting *Red* becomes more attractive and lies are strategic complements.

The experiment can be thought of as a more simplified and decomposed version of AN&R's second experiment on lying and image concerns. As in their game, participants with different material incentives to lie decide whether to lie or not. When lying, they pool with participants from all other states. The current experiment simplifies the game by focusing on three different payoff states instead of ten, which should be sufficient to generate differences in the quality of liars. It also gets at decomposing credibility and quality by (i) either introducing an experimental manipulation which changes the credibility of reporting the highest state, holding constant the composition of liars, or, (ii) introducing an experimental manipulation which changes the composition of liars while having only a second-order effect on the credibility of reporting the highest state.

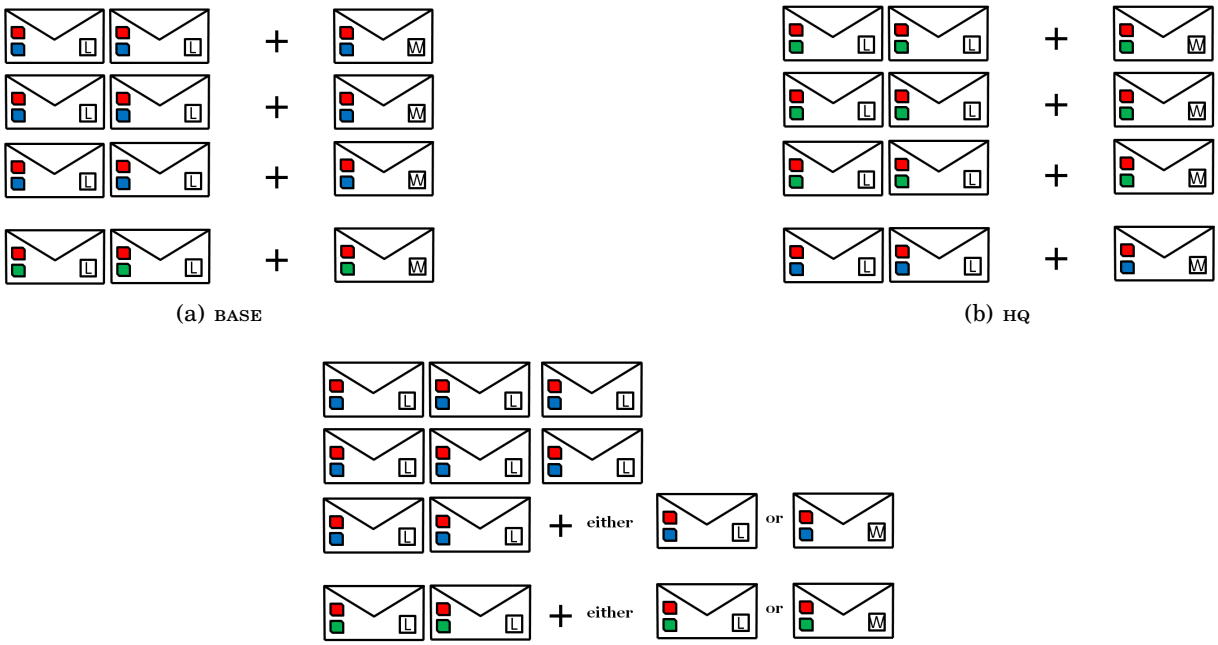
4.1.2 Implementation

The experimental lying game follows the code-reporting design introduced by Feess and Kerzenmacher (2018). Participants draw a random envelope from a box. The envelope consists of two parts, a code sheet and a lottery ticket. Each envelope contains either a *Red-Blue* or a *Red-Green* code sheet. *Red-Blue* code sheets include two four-digit codes for *Red* and *Blue* and *Red-Green* code sheets include two four-digit codes for *Red* and *Green*. The lottery ticket can either be a win or a loss. Upon finding a win in the envelope, participants are instructed to write down the code for *Red* on a sheet of paper. In case of a loss they are told to write down the code for the other color. After all participants write down one code, they throw the envelope with the code sheet and lottery ticket in a box. Thereafter, they type their code into the computer, which then registers the lottery payoff. The codes for each color are the same for all participants, so that it is impossible for the experimenter to find out whether participants really typed the code into the computer that they were supposed to. The procedure gives participants a clear monetary incentive to lie and to report the code for *Red* instead of *Blue* or *Green*. At the same time, in case they are going to lie, participants have to lie to the full extent and report *Red*.

4.1.3 Treatments and hypotheses

There are three treatments which vary the probabilities of drawing *Red*, *Blue*, or *Green*. Figure 6 displays the distribution of envelopes in each of them. In all treatments, twelve participants take part in one session at the same time. In BASE, four participants in each session are instructed to report *Red* (which pays 12 euros), six to report *Blue* (which pays 10 euros), and two to report *Green* (which pays 2 euros). In the high quality treatment, HQ, the proportion of participants drawing *Blue* and *Green* switches compared to baseline. Thus, while the average liar draws *Blue* in BASE, the average liar will draw *Green* in HQ. A participant who draws *Green* is

Figure 6. Treatments



(a) BASE (b) HQ

(c) LP

Note: All envelopes contain codes for *Red* and one additional color, as indicated by the blue and green dots. Envelopes with an "L" contain a loss and envelopes with a "W" contain a win.

relatively unlucky and would earn only a small amount. Therefore, lying after drawing *Green* is arguably more justified than lying after drawing *Blue*. In the language of the model, liars in HQ on average are of higher quality than liars in BASE. Comparing lying behavior between BASE and HQ gives an indication of how the composition of liars influences lying behavior. As more participants draw *Green* in HQ, the average liar has a better reason to lie. The positive image spillovers on other participants will increase lying, conditional on the draw.

Hypothesis 1 (Related to proposition 4)**.** *The fraction of participants lying, conditional on the draw, is higher in HQ than in BASE.*

In the low probability treatment, LP, only one participant in each session draws *Red* as compared to four in BASE. Conversely, more participants draw one of the lower states The winning lot for the LP treatment was included in one random envelope for each session before the experiment.[16] Therefore, reporting *Red* is less credible in LP, which should lead to a decrease in conditional lying.

Hypothesis 2 (Related to proposition 3)**.** *The fraction of participants lying, conditional on the draw, is lower in LP than in BASE.*

[16]Because of this, and in contrast to the remaining treatments, the draw distribution in LP is not deterministic. This was mainly done because it has the advantage of substantially reducing the probability of drawing *Red* as compared to BASE, while at the same time ensuring a positive probability of winning after drawing either *Red-Blue* or *Red-Green*. Otherwise, motives such as spite and experimenter demand effects might motivate some participants in LP to lie if they draw a code combination they never could have won with anyway.

4.1.4 Additional experimental features

Additional experimental features are summarized below. The precise instructions of the experiment can be found in appendix D.

Before the lying game and after reading the experimental instructions, participants have to state their beliefs about how many participants they think reported *Red*, *Blue* and *Green* in a previous session of the same game. Participants are incentivized to report their true beliefs with a quadratic scoring rule.[17] The timing of the belief elicitation before the choice is in line with Rauhut (2013), who runs a lying game with and without belief elicitation prior to choice and finds no impact of the elicitation on behavior.

To increase the image aspect, all participants have to stand up at the end of the experiment and announce the color that they reported in front of the experimenter and all other participants. This is known to the participants before they take their reporting decision. Because participants draw from a lottery without replacement, during the public announcements participants can find out about the exact number of liars, ruling out chance as an alternative explanation.

One might worry that a lottery without replacement increases the perceived fear of punishment among participants. Participants could for example fear individual or collective punishment if the report and draw distribution in their session do not coincide. Because of this, the instructions are careful in describing to participants that their draws can not be tracked individually, that they will receive the money corresponding to their report, and that there will not be an additional stage of the experiment after they make their public announcements.

4.2 Procedures

The experimental sessions were run in November and December 2019 in the WZB/TU lab in Berlin. Recruitment was done via ORSEE (Greiner, 2015) and the experimental software was programmed with oTree (Chen, Schonger, and Wickens, 2016). Participants first read the instructions in private and after everyone finished they were verbally summarized by the experimenter. Thereafter, participants had to answer control questions that tested their understanding of different aspects of the game. After the belief elicitation and lottery decision, participants answered a number of demographic questions and an open question about how they took the decision in the experiment. Sessions lasted for around 30 minutes. In addition to their experimental earnings, participants received 5€ as a show-up fee. Average earnings were around 18€. In total 360 participants took part in the experiment (42% female), 144 each in BASE and in HQ and 72 in LP. There are less observations in LP because only one participant draws *Red* in LP compared to four in the other treatments, i.e. more participants take an active decision in each session of LP. The preregistration of the experimental game, of the main hypotheses and their analysis, and of the sample size is available under

[17] For each participant the computer randomly chooses their stated belief for one of the three states as payoff relevant. Participants then have to pay $4€ - 0.049€ \times (\#\text{Participants who reported the color in previous session} - \#\text{Participant's estimate})^2$.

https://aspredicted.org/blind.php?x=ec77bs.

5 Results

This section will first examine choice behavior before going over to beliefs.

5.1 Choice behavior

Figure 7 presents results. The table in the left panel shows regressions of an indicator whether a participant lied or not on a treatment dummy, controlling for the draw.[18] There are two major takeaways from the table. First, monetary stakes significantly impact lying. Participants who draw *Blue*, and thus have a smaller monetary gain from lying, are around 27 percentage points less likely to lie than participants who draw *Green*. The right panel in figure 7 further shows that this holds individually in all treatments. Second, the regressions show no evidence that signaling concerns matters for lying. When comparing lying behavior between BASE and HQ in column 1, there is a small but insignificant treatment effect. Column 2 compares behavior between BASE and LP and finds a nonsignificant treatment effect with a point estimate very close to zero.

Result 1 (Related to hypothesis 1)**.** *The fraction of participants lying, conditional on the draw, is not significantly different between HQ and BASE.*

Result 2 (Related to hypothesis 2)**.** *The fraction of participants lying, conditional on the draw, is not significantly different between LP and BASE.*

5.2 Beliefs

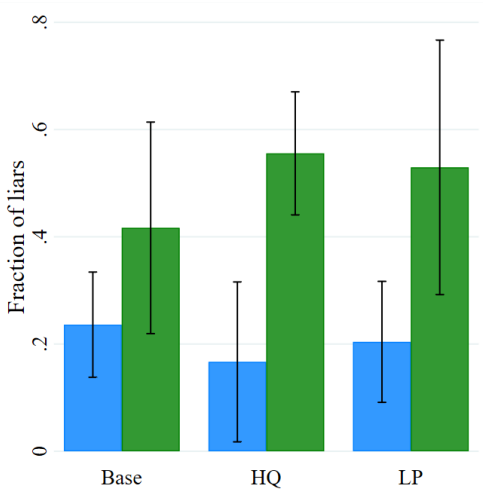
The null results from the previous part could be driven by two distinct reasons. Either, image concerns at most had a small effect on behavior in this experiment, or participants are image concerned but their beliefs are inconsistent with equilibrium strategies.

In fact, beliefs are very heterogenous within treatments. Figure 8 shows scatter plots for stated beliefs about how many participants reported *Green* and *Blue*. The figure also highlights two kinds of stated beliefs that are inconsistent with basic image models of lying. Beliefs that fall in the region to the left of the dashed line and below the shaded area imply that the fraction of *Blue* participants lying is higher than the fraction of *Green* participants lying. This violates the property that more individuals lie the higher the monetary gain from lying. Beliefs in the shaded area are only consistent with downwards lying from *Red* participants. The majority of stated beliefs are however free of violations, with 76-90% of participants in each treatment stating beliefs that are consistent with both properties.

[18] Lying cannot observed individually, but because the draw distribution is deterministic, the individual data used in the regression can be recovered.

Figure 7. Choice behavior and treatment effects

	(1)	(2)
Blue	−0.278***	−0.260***
	(0.075)	(0.087)
HQ	0.028	
	(0.075)	
LP		0.001
		(0.070)
Constant	0.490***	0.476***
	(0.076)	(0.083)
Observations	192	162
R^2	0.092	0.064



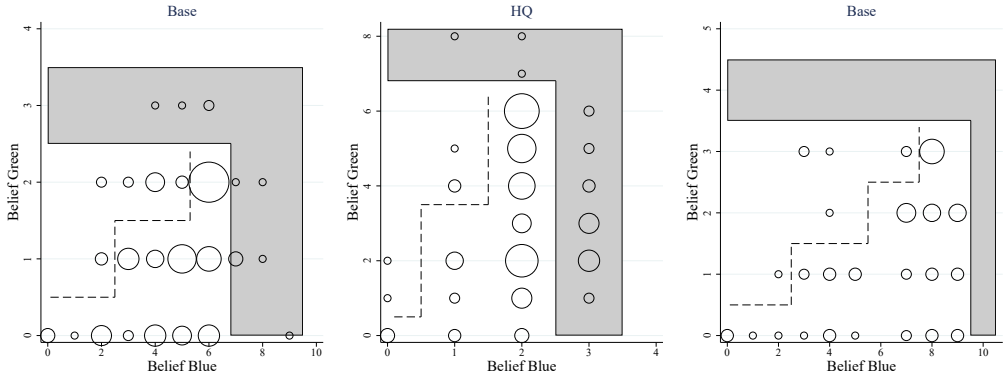
Note: The left panel presents OLS regressions with lie as dependent variable. *Blue* is a dummy equal to one if the participant drew *Blue*. HQ and LP are two treatment dummies. Column 1 presents regressions using only observations from treatments BASE and HQ. Column 2 presents regressions using only observations from treatments BASE and LP. All regressions exclude observations of participants that drew *Red*. Robust standard errors in parenthesis. *** $p < 0.01$. The right panel displays the fraction of liars who drew *Blue* and *Green* by treatment. Error bars display 95% confidence intervals.

5.2.1 Coding liars

Two additional lab sessions were run to obtain more precise data on the heterogeneity of beliefs. At the end of the main experiment, participants answered an open question that asked them to reason about how they took their decision in the experiment. Only participants who did not take part in the original experiment took part in the additional sessions. They worked as evaluators and were tasked to read the answers to the open question in the original experiment and to judge which of the participants they suspect of lying. Table 1 contains explanations from six participants reporting *Red* in one example session of BASE. Since only four participants drew *Red*, but six participants reported *Red*, two of these four participants must have lied.[19]

[19]Downwards lying was in principle possible in the experiment and thus it could be the case that more than two out of the six participants in the table lied. This is however very unlikely. *Red* was never reported by less than the number of participants who drew it in any of the 33 experimental sessions. In addition, experimental evidence on downwards lying is only observed for very selected samples and special design features. In an experiment with nuns, Utikal and Fischbacher (2013) find evidence that is consistent with downward lying. Barron (2019) finds more systematic evidence that lab participants lie downwards on a small stakes die when they simultaneously have the opportunity to lie upwards on a high stakes die. Throughout the results discussion I will maintain the assumption that no participant lied downwards.

Figure 8. Heterogeneity of beliefs



Note: Beliefs about the number of participants reporting *Blue* and *Green*. Larger bubbles indicate more observations. Stated beliefs above the dashed line and below the shaded area violate the monotonicity property that individuals lie more if stakes increase. Stated beliefs in the gray shaded area violate the no downwards lying condition.

Table 1. Example answers

Participant id	Reasoning	Coded suspicion	$\hat{lie}$
1	I am here to earn as much money as possible in the shortest time.	0.921	1
2	I wanted to earn a lot of money.	0.868	1
3	I took all of my decisions in accordance with the rules.	0.079	0
4	By drawing the main prize	0.053	0
5	Honestly	0.026	0
6	I was the last one and took the envelope	0.053	0

Note: The table reports answers of participants reporting *Red* to the question *"Please give a concise explanation of how you took your decision in the experiment"* in one session of BASE. Answers and the wording of the question have been translated from German.

Evaluators in the coding sessions sequentially read all explanations participants reporting *Red* gave in the main experiment and had to evaluate whether a participant reported *Red* dishonestly or not. They saw answers of all participants reporting *Red* in the same session at the same time, as in table 1. The sequence of sessions and of answers within sessions were presented to each evaluator in random order. These additional sessions lasted for one hour each. Evaluators earned a flat wage of 10 euros and could gain 7 additional euros if they judged one randomly chosen answer in the same way as a random evaluator in the same session. This additional incentive was given mainly to keep evaluators motivated to read and evaluate each answer carefully. In total 38 participants took part in the additional sessions.

The evaluators' assessments were largely consensual. They coded each individual answer in the direction of the modal evaluation in 85% of all cases, an improvement over the case where everyone answered randomly, which would have coincided with the modal answer only 65% of the time.

From the coded evaluations I generate a suspicion indicator which is the fraction of evalua-

tors that coded the participant as dishonest. The indicator $\hat{lie}$ classifies a participant as a liar if they were amongst the participants who gave the most suspicious answers in a session. In the example in table 1, participants 1 and 2 get assigned a $\hat{lie}$ of one because two participants in that sessions must have lied and they got the highest coded suspicion of all participants in the same session.

5.2.2 Signaling motives and beliefs

To investigate whether individual-level behavior and beliefs are consistent with image concerns, I create two indicators capture credibility and quality beliefs. The belief about the proportion of participants who tell the truth, conditional on reporting *Red*, is taken as an indicator for credibility. The belief about the proportion liars who drew *Green* is an indicator of the quality of liars.

Figure 9 presents average credibility and quality beliefs, by action and treatment. Comparing BASE with HQ, the belief about quality of liars is significantly higher in HQ ($p < 0.001$, Mann-Whitney test). Similarly, moving from BASE to LP, the credibility belief decreases ($p < 0.001$, Mann-Whitney test). These results suggest that, while the treatments were effective in shifting participant beliefs, these shifts did not affect behavior.

There is further evidence going against signaling motives. Compare beliefs of nonliars between BASE and HQ. While the average credibility belief of nonliars is similar, average quality beliefs are higher among nonliars in HQ (credibility: 0.781 BASE, 0.775 HQ, $p = 0.817$, Mann-Whitney test; quality: 0.384 BASE, 0.605 HQ, $p = 0.017$, Mann-Whitney test). That is, nonliars in HQ hold beliefs consistent with a higher image of reporting *Red* than nonliars in BASE. If participants are motivated by high image concerns, credibility and quality beliefs of non-liars should be the same *conditional* on telling the truth. A similar point can be made when comparing nonliars between HQ and LP. Nonliars in LP have lower credibility beliefs, while quality beliefs are similar (credibility: 0.524 LP, 0.775 HQ, $p < 0.001$, Mann-Whitney test; quality: 0.605 LP, 0.516 HQ, $p = 0.324$, Mann-Whitney test).[20]

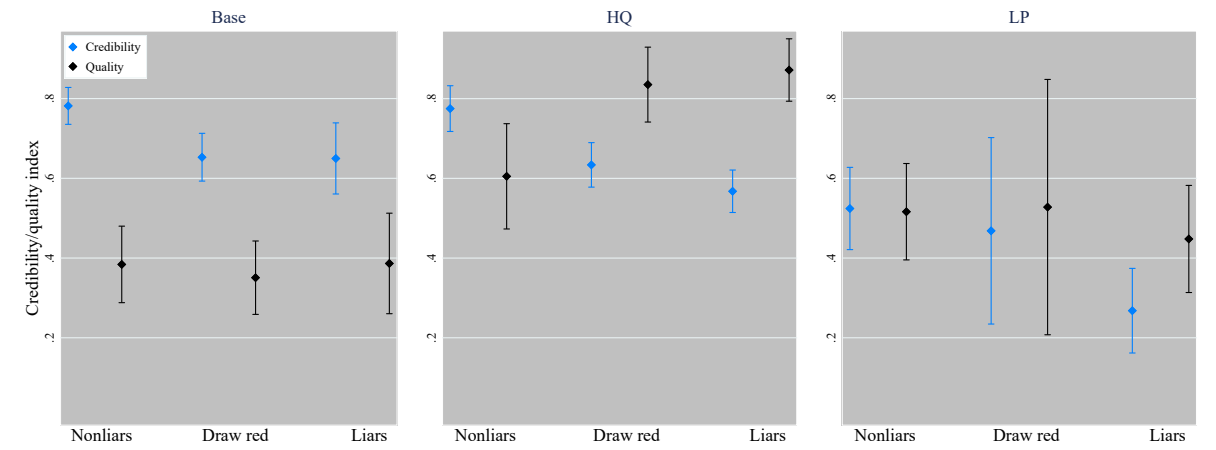
5.2.3 Relationship between beliefs and lying behavior

If image concerns cannot explain the data, the question arises if beliefs have any causal role in determining behavior. One suggestive piece of evidence is that participants might be conformists; within every treatment, liars think that more participants lie than nonliars ($p < 0.006$, Mann-Whitney tests).[21] This correlation is documented in other studies that investigate lying and beliefs about lying (Abeler, Becker, and Falk, 2014, Hugh-Jones, 2016, Abeler et al., 2019). AN&R caution against a causal interpretation, because priors and preferences are likely correlated. This can for example happen because of consensus effects, where individuals overweight

[20] This argument ignores that nonliars on average have higher financial incentives to lie in HQ. Incorporating financial incentives would only lead to the stronger claim that credibility and quality beliefs of nonliars should actually be lower in HQ.

[21] Testing instead whether participants reporting *Red* hold higher beliefs than other participants gives similar results ($p < 0.013$ Mann-Whitney tests).

Figure 9. Beliefs of liars, nonliars, and participants who drew *Red*



Note: Nonliars are participants who reported *Blue* or *Green*. Participants who reported *Red* are classified as either lying or to have drawn red based on the $\widehat{\text{lie}}$ indicator from the coding sessions. Error bars are 95% confidence intervals.

their own preferences when forming beliefs (see discussions in Lundquist, Ellingsen, Gribbe, and Johannesson, 2009, Blanco, Engelmann, Koch, and Normann, 2014). This part investigates whether the observed association could be causal.

Consider a model of conformity, where participants incur moral costs of lying which decrease if the participant believes that many others are lying. Formally, assume a linearly additive utility function of the form

$$u(j,t,b,a) = y(a) + 1_{a\neq j}(\theta b - t), \tag{12}$$

where b is the participant's belief about the fraction of participants who are lying and θ is a sensitivity parameter. The notation is otherwise unchanged.

If we are to estimate such a model using the data from the experiment, we have to overcome two challenges; (i) the actions (lying /truth-telling) are not directly observed and (ii) the consensus effect discussed above gives reasons to assume a confounding, non-causal relationship between preferences and beliefs. In face of these challenges, I develop and estimate a structural model of lying that explicitly models the classification error and allows for a potential correlation between beliefs and preferences.

We assume that participants have lying costs which are drawn from a log-normal distribution where log-costs have mean m and standard deviation σ_m. Each participant's stated belief is drawn from a normal distribution which, since beliefs are restricted to lie between zero and one, is censored at the end points. Beliefs further depend on the treatment and can potentially be correlated with the lying cost. The equation for a stated belief b of a participant becomes

$$b^* = \beta_{\text{treatment}} + \rho t + \varepsilon, b = \begin{cases} 0 & \text{if } b^* < 0 \\ b^* & \text{if } b^* \in [0,1] \\ 1 & \text{if } b^* > 1, \end{cases}$$

where $\beta_{\text{treatment}}$ is a treatment indicator, t is the (unobserved) lying cost, ρ is a parameter that measures the correlation between lying costs and beliefs and ε is a normally distributed i.i.d. error term with expectation zero and standard deviation σ_b. The modeling of the belief function closely follows the empirical framework of Bellemare, Sebald, and Strobel (2011), who control for a potential correlation between preferences and beliefs when estimating guilt aversion. A negative ρ is consistent with a consensus effect, as it implies that participants with lower lying costs believe that a large fraction of participants are lying. Intuitively ρ is identified across treatments as the treatments generate variation in beliefs which should be uncorrelated with intrinsic preferences. Therefore, between-treatment variations in behavior can be attributed to a direct causal impact of beliefs on lying, while the remaining within-treatment variation in beliefs and behavior will be attributed to the consensus effect.

Assume that the utility derived from lying and truth-telling is further influenced by an i.i.d. error term which follows the extreme-value type 1 distribution, so that a standard logit choice model arises, where the probability of a participant lying after drawing *Green* is equal to

$$P(\text{report } Red | \text{draw } Green, b_i, t_i) = \frac{\exp\{u(Green, t_i, b_i, Red)\}}{\exp\{u(Green, t_i, b_i, Red)\} + \exp\{u(Green, t_i, b_i, Green)\}},$$

and analogously for a participant who drew *Blue*. We further maintain the assumption of no downwards lying. This implies that a participant who drew *Red* always reports *Red* with probability one.

I estimate the model via maximum likelihood. One challenge when estimating the model is that we do not perfectly observe lies, as participants who reported *Red* could have either reported truthfully or lied. This is a problem of *classification error*, which can be dealt with by explicitly taking account of the probability of misclassification in the likelihood function (Hausman, Abrevaya, and Scott-Morton, 1998). Appendix C includes details on the estimation.

Table 2 presents estimates, first for two restricted models which ignore the relationship between beliefs and lying and between preferences and beliefs. The model in the last column takes account of both. It is interesting to observe that the LC + Beliefs model, which ignores the correlation between beliefs and preferences, predicts a strongly positive relationship between beliefs about lying and behavior. Its estimates imply that a ten percentage point increase in the belief among participants leads to a seven percentage point increase in the fraction of participants who lie. This relationship becomes small and insignificant once we allow for a correlation between beliefs and preferences in the fully specified model. The correlation in turn is significantly negative, consistent with a consensus effect, where participants with a higher wilingness to lie also expect many others to lie.

6 Discussion

The paper presented a model where agents derive reputational esteem if they are perceived as being of honest character. Such a model can explain many of the previous experimental results on lying games. While it was narrowly applied to study games of the die-roll type, the model

Table 2. Likelihood estimates

	LC only	LC + Beliefs	Fully specified
m	2.209***	2.621***	2.363***
	(0.163)	(0.1)	(0.169)
σ_m	1.232***	0.332***	1.051***
	(0.442)	(0.089)	(0.355)
θ		16.168***	3.337
		(3.472)	(2.634)
ρ			−0.021**
			(0.009)
β_{Base}	0.197***	0.197***	0.51***
	(0.027)	(0.029)	(0.076)
β_{HQ}	0.278***	0.278***	0.577***
	(0.027)	(0.028)	(0.08)
β_{LP}	0.199***	0.199***	0.495***
	(0.039)	(0.039)	(0.085)
σ_b	0.31***	0.31***	0.212***
	(0.014)	(0.015)	(0.017)
Log-likelihood	−350.012	−327.262	−318.578

Note: Asymptotic standard errors in parenthesis. ** $p < 0.05$, *** $p < 0.01$.

could be modified to study relevant economic environments, such as credence goods markets or markets for financial advice. The paper aimed to clarify some of the differences between the model and other theoretical approaches which conceptualize the reputational cost of lying by interpreting a lie as a bad deed. The predictions between both approaches mainly differ because the signaling motive in the character-based image model is driven by an additional quality effect.

The experimental results do not show evidence for belief-dependent preferences. They fail to confirm a new quality effect which is theoretically explored in this paper, but also find no evidence that credibility of the report matters. The results therefore are largely in contrast with previous experiments that find support for the credibility effect (Feess and Kerzenmacher, 2018, Gneezy et al., 2018, Abeler et al., 2019).

One important aspect that could have influenced the experimental results in this paper is that the instructions were relatively clear that lies would not be punished and that there was no trick involved in the experiment.[22] Perceived fear of punishment has been shown to influence lying behavior in the standard game with replacement (Kajackaite and Gneezy, 2017). The threat of punishment can provide instrumental reasons to appear credible if there is a risk of being caught cheating and punished. Therefore, one reason for the observed differences might be that some of the behavior consistent with image concerns in previous experiments was trig-

[22]Some evidence that participants understood this comes from the answers to the open question at the end of the experiment. The majority of truth-tellers, for example, justifies their behavior on the moral ground that lying is bad, instead of fear of punishment.

gered by a fear of punishment and does not constitute a pure (non-instrumental) image concern. This interpretation is also consistent with the experimental finding that belief-dependent preferences do not seem to matter in sender-receiver games (López-Pérez and Spiegelman, 2013), where ambiguity about the game and about punishment capabilities of the receiver should be much less pronounced. López-Pérez and Spiegelman set up a sender-receiver game and, similarly to this experiment, vary the distribution of draws. Like in the present experiment, they find that this manipulation does not affect behavior, even though they observe a strongly positive correlation between behavior and beliefs. The present experiment has however little to say about whether perceived punishment motivates the type of disguisive behavior typically observed in experimental lying games. One promising direction for future research would be to more systematically test its effect.

The current experiment also differs from former experiments in that the audience participants signaled to was extended. In standard lying experiments, participants usually only report to the experimenter, without an additional public announcement. This might have had the unintended consequence that participants were signaling to multiple audiences, which might obscure their signaling motive. Public reports of actions are a usual way of intensifying image concerns in experiments (Ariely, Bracha, and Meier, 2009, Bursztyn and Jensen, 2017, Friedrichsen and Engelmann, 2018). This might however not always trigger the intended signaling motive. A striking example of this is Dufwenberg and Muren (2006) who conduct a dictator game experiment with first-year economics students. The dictator giving rates are either kept anonymous or publicly announced. Dufwenberg and Muren find that students actually become less generous if their action is made public. While these concerns cannot be ruled out, the current experiment finds little evidence that beliefs have any causal effect on behavior.

The experimental results do however confirm previous evidence on the consensus effect in the lying domain. In particular, I provide quantitative evidence that beliefs about lying are correlated with behavior because preferences and beliefs are correlated. It would be interesting to theoretically think through formal concepts that can accommodate consensus effects and study their strategic implications. From an empirical perspective, the framework that I present explicitly takes account of the classification error inherent in experimental lying games. The model provides consistent parameter estimates and could be used in future research to empirically estimate relations between individual-level characteristics and actions in lying games.

References

Abeler, J., A. Becker, and A. Falk (2014): "Representative Evidence on Lying Costs," *Journal of Public Economics*, 113, 96–104.

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 1115–1153.

Akerlof, G. A. (1970): "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 84, 488.

Ariely, D., A. Bracha, and S. Meier (2009): "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *The American Economic Review*, 99, 544–555.

Bagnoli, M. and T. Bergstrom (2005): "Log-Concave Probability and Its Applications," *Economic Theory*, 26, 445–469.

Barron, K. (2019): "Lying to Appear Honest," *WZB Discussion Paper SP II 2019-307*.

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

——— (forthcoming): "Belief-Dependent Motivations and Psychological Game Theory," *Journal of Economic Literature*.

Bellemare, C., A. Sebald, and M. Strobel (2011): "Measuring the Willingness to Pay to Avoid Guilt: Estimation Using Equilibrium and Stated Belief Models," *Journal of Applied Econometrics*, 26, 437–453.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

——— (2011): "Laws and Norms," *NBER Working Paper*.

Blanco, M., D. Engelmann, A. K. Koch, and H. T. Normann (2014): "Preferences and Beliefs in a Sequential Social Dilemma: A within-Subjects Analysis," *Games and Economic Behavior*, 87, 122–135.

Bursztyn, L. and R. Jensen (2017): "Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure," *Annual Review of Economics*, 9, 131–153.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Cohn, A., M. A. Maréchal, D. Tannenbaum, and C. L. Zünd (2019): "Civic Honesty around the Globe," *Science*, 365, 70–73.

Crawford, V. P. and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431.

Diekmann, A., W. Przepiorka, and H. Rauhut (2015): "Lifting the Veil of Ignorance: An Experiment on the Contagiousness of Norm Violations," *Rationality and Society*, 27, 309–333.

Dufwenberg, M. and M. A. Dufwenberg (2018): "Lies in Disguise – A Theoretical Analysis of Cheating," *Journal of Economic Theory*, 175, 248–264.

Dufwenberg, M. and A. Muren (2006): "Generosity, Anonymity, Gender," *Journal of Economic Behavior & Organization*, 61, 42–49.

Feess, E. and F. Kerzenmacher (2018): "Lying Opportunities and Incentives to Lie: Reference Dependence versus Reputation," *Games and Economic Behavior*, 111, 274–288.

Fischbacher, U. and F. Föllmi-Heusi (2013): "Lies in Disguise-an Experimental Study on Cheating," *Journal of the European Economic Association*, 11, 525–547.

Friedrichsen, J. and D. Engelmann (2018): "Who Cares about Social Image?" *European Economic Review*, 110, 61–77.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1, 60–79.

Gibson, R., C. Tanner, and A. F. Wagner (2013): "Preferences for Truthfulness: Heterogeneity among and within Individuals," *American Economic Review*, 103, 532–548.

Gneezy, U., A. Kajackaite, and J. Sobel (2018): "Lying Aversion and the Size of the Lie," *American Economic Review*, 108, 419–453.

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013): "Measuring Lying Aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

Grant, A. (2019): "Tweet, https://twitter.com/AdamMGrant/status/1170320617998082048," .

Greiner, B. (2015): "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

Grossman, Z. and J. J. van der Weele (2017): "Self-Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association*, 15.

Hausman, J., J. Abrevaya, and F. Scott-Morton (1998): "Misclassification of the Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics*, 87, 239–269.

Hugh-Jones, D. (2016): "Honesty, Beliefs about Honesty, and Economic Growth in 15 Countries," *Journal of Economic Behavior & Organization*, 127, 99–114.

Hursthouse, R. and G. Pettigrove (2018): "Virtue Ethics," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, vol. Winter 2018 Edition.

Kajackaite, A. and U. Gneezy (2017): "Incentives and Cheating," *Games and Economic Behavior*, 102, 433–444.

Kartik, N. (2009): "Strategic Communication with Lying Costs," *Review of Economic Studies*, 76, 1359–1395.

Khalmetski, K. and D. Sliwka (2019): "Disguising Lies—Image Concerns and Partial Lying in Cheating Games," *American Economic Journal: Microeconomics*, 11, 79–110.

López-Pérez, R. and E. Spiegelman (2013): "Why Do People Tell the Truth? Experimental Evidence for Pure Lie Aversion," *Experimental Economics*, 16, 233–247.

Lundquist, T., T. Ellingsen, E. Gribbe, and M. Johannesson (2009): "The Aversion to Lying," *Journal of Economic Behavior & Organization*, 70, 81–92.

Peterson, J. (2018): *12 Rules for Life: An Antidode to Chaos*, Penguin Random House.

Rauhut, H. (2013): "Beliefs about Lying and Spreading of Dishonesty: Undetected Lies and Their Constructive and Destructive Social Dynamics in Dice Experiments," *PLoS ONE*, 8, e77878.

Ruffle, B. J. and Y. Tobol (2017): "Clever Enough to Tell the Truth," *Experimental Economics*, 20, 130–155.

Utikal, V. and U. Fischbacher (2013): "Disadvantageous Lies in Individual Decisions," *Journal of Economic Behavior & Organization*, 85, 108–111.

A Proofs

Notation Throughout the proof section, I will use $\mathcal{M}(\hat{t}) \equiv E(t|t \leq \hat{t})$ to simplify notation.

A.1 Proof of proposition 1

We first provide two lemmatas before proceeding with the proof.

Lemma 1 (Properties of $\hat{t}(\Delta(K,j),\varphi)$)**.** *The following properties hold for $\hat{t}(\Delta(K,j),\varphi)$ if $j \notin \mathcal{Q}$ and μ is small enough, $E(t|t>\hat{t})$ is convex in $\hat{t}$.*

(i) $\frac{\mathrm{d}\hat{t}(\Delta(K,j),\varphi)}{\mathrm{d}\varphi} \in (0,1)$.

(ii) $\frac{\mathrm{d}^2\hat{t}(\Delta(K,j),\varphi)}{\mathrm{d}\varphi \mathrm{d}\Delta(K,j)} \leq 0$.

Proof. $\hat{t}(\Delta(K,j),\varphi)$ is implicitly defined in

$$\hat{t} + \mu\left[\mathcal{R}_j(\hat{t}) - \varphi\right] - \Delta(K,j) = 0.$$

(i) Implicitly differentiating the equation brings

$$\frac{\mathrm{d}\hat{t}(\Delta(K,j),\varphi)}{\mathrm{d}\varphi} = \frac{\mu}{1+\mu\mathcal{R}_j'(\hat{t}(\Delta(K,j),\varphi))} \text{ if } j \leq k^*,$$

$\mathcal{R}_j'(\hat{t}(\Delta(K,j),\varphi)) = \frac{\mathrm{d}E(t|t>\hat{t}(\Delta(K,j),\varphi))}{\mathrm{d}\hat{t}} > 0$. Therefore, the derivative is between 0 and 1 if μ is small (e.g. $\mu \leq 1$).

(ii) The cross-derivative is

$$\frac{\mathrm{d}^2\hat{t}(\Delta(K,j),\varphi)}{\mathrm{d}\varphi \mathrm{d}\Delta(K,j)} = -\frac{\mu\mathcal{R}_j''(\hat{t}(\Delta(K,j),\varphi))\frac{\mathrm{d}\hat{t}(\Delta(K,j),\varphi)}{\mathrm{d}\Delta(K,j)}}{(1+\mu\mathcal{R}_j'(\hat{t}(\Delta(K,j),\varphi)))^2} \leq 0,$$

as it is easily verified that $\frac{\mathrm{d}\hat{t}(\Delta(K,j),\varphi)}{\mathrm{d}\Delta(K,j)} > 0$ and, as $\mathcal{R}_j(\hat{t}) = E(t|t>\hat{t})$, $\mathcal{R}_j''(\hat{t}) \geq 0$. ■

Lemma 2 (Properties of $\mathcal{L}(\varphi)$)**.** *$\mathcal{L}(\varphi)$ is (i) a continuous function with (ii) $\mathcal{L}'(\varphi) < 1$. There exists (iii) a unique value $\xi \in (0, E(t))$ such that $\mathcal{L}(\xi) = \xi$.*

Proof. (i) The functions $\hat{t}_j(\varphi)$, $F(t)$ and $\mathcal{M}(t)$ are continuous functions. It follows that $\mathcal{L}(\varphi)$ is continuous on each range of values for φ that do not change the threshold state k^*.
To see that $\mathcal{L}(\varphi)$ is continuous around the values of φ that induce a change in k^*, we examine the case where $y(K)+\mu\varphi' = y(k')+\mu E(t)$. Consider an $\varepsilon > 0$ which is arbitrarily small. Obviously, permuting φ' to $\varphi'+\varepsilon$ induces a continuous change in $\mathcal{L}(\varphi)$ as it does not change k^*. If we look

at $\varphi' - \varepsilon$, the threshold state changes to $k' - 1$ and thus

$$\begin{aligned}
\mathcal{L}(\varphi' - \varepsilon) &= \frac{\sum_{j=1}^{k'-1} F(\hat{t}_j(\varphi' - \varepsilon))\mathcal{M}(\hat{t}_j(\varphi' - \varepsilon))}{\sum_{j=1}^{k'-1} F(\hat{t}_j(\varphi' - \varepsilon))} \\
&= \frac{\sum_{j=1}^{k'} F(\hat{t}_j(\varphi' - \varepsilon))\mathcal{M}(\hat{t}_j(\varphi' - \varepsilon))}{\sum_{j=1}^{k'-1} F(\hat{t}_j(\varphi' - \varepsilon))} - \frac{F(\hat{t}_{k'}(\varphi' - \varepsilon))\mathcal{M}(\hat{t}_{k'}(\varphi' - \varepsilon))}{\sum_{j=1}^{k'-1} F(\hat{t}_j(\varphi' - \varepsilon))} \\
&= \underbrace{\frac{\sum_{j=1}^{k'} F(\hat{t}_j(\varphi' - \varepsilon))}{\sum_{j=1}^{k'-1} F(\hat{t}_j(\varphi' - \varepsilon))}}_{a(\varphi' - \varepsilon)} \left[\underbrace{\frac{\sum_{j=1}^{k'} F(\hat{t}_j(\varphi' - \varepsilon))\mathcal{M}(\hat{t}_j(\varphi' - \varepsilon))}{\sum_{j=1}^{k'} F(\hat{t}_j(\varphi' - \varepsilon))}}_{b(\varphi' - \varepsilon)} - \underbrace{\frac{F(\hat{t}_{k'}(\varphi' - \varepsilon))\mathcal{M}(\hat{t}_{k'}(\varphi' - \varepsilon))}{\sum_{j=1}^{k'} F(\hat{t}_j(\varphi' - \varepsilon))}}_{c(\varphi' - \varepsilon)} \right]
\end{aligned}$$

By continuity of $\hat{t}_j(\varphi)$ for $j < k'$, and by continuity of $F(t)$ it follows that $\lim_{\varepsilon \to 0} F(\hat{t}_j(\varphi' - \varepsilon)) = F(\hat{t}_j(\varphi'))$ for $j < k'$. Furthermore, since for $\varphi' - \varepsilon$ the threshold state is $k' - 1$, by definition of the threshold function, $\hat{t}_{k'}(\varphi' - \varepsilon) = 0$.
An increase in the threshold state is followed by a continuous change of the threshold function. To see this, consider the threshold function $\hat{t}_{k'}(\varphi')$, which is implicitly defined in

$$\hat{t}_{k'} + \mu \left[\mathcal{R}_{k'}(\hat{t}_{k'}) - \varphi' \right] - \Delta(K, k') = 0.$$

Plugging in $\varphi' = E(t) - \frac{\Delta(K,k')}{\mu}$, it becomes visible that the only solution for $\hat{t}_{k'}(\varphi') = 0$. It follows that $\lim_{\varepsilon \to 0} \hat{t}_{k'}(\varphi' - \varepsilon) = 0$ and therefore $\lim_{\varepsilon \to 0} F(\hat{t}_{k'}(\varphi' - \varepsilon)) = F(0) = 0$. This implies that $\lim_{\varepsilon \to 0} a(\varphi' - \varepsilon) = 1$. Moreover, $\mathcal{M}(0) = 0$ and therefore $\lim_{\varepsilon \to 0} c(\varphi' - \varepsilon) = 0$. It follows that

$$\lim_{\varepsilon \to 0} \mathcal{L}(\varphi' - \varepsilon) = \lim_{\varepsilon \to 0} b(\varphi' - \varepsilon) = \frac{\sum_{j=1}^{k'} F(\hat{t}_j(\varphi))\mathcal{M}(\hat{t}_j(\varphi'))}{\sum_{j=1}^{k'} F(\hat{t}_j(\varphi'))} = \mathcal{L}(\varphi').$$

The function is thus continuous.

(*ii*) Taking the derivative of (3) with respect to φ yields

$$\mathcal{L}'(\varphi) = \sum_{j=1}^{k^*} \mathrm{P}(\text{draw } j|\text{lie}) \mathcal{M}'(\hat{t}_j(\varphi)) \frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} + \sum_{j=1}^{k^*} \frac{\mathrm{dP}(\text{draw } j|\text{lie})}{\mathrm{d}\varphi} \mathcal{M}(\hat{t}_j(\varphi)) \tag{13}$$

We first show that the first term is smaller than one. Consider the term $t - \mathcal{M}(t)$, which can be rewritten as $\frac{\int_0^t F(s)\,\mathrm{d}s}{F(t)}$. By log-concavity of $f(t)$ the integral in the fraction is log-concave, which implies that $t - \mathcal{M}(t)$ is increasing.[23] From this, we conclude that $\mathcal{M}'(t) \in (0,1)$. Further, $\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} \in$

[23] This follows from the property that for a log-concave function $g(x)$, $\frac{g(x)}{g'(x)}$ is increasing in x. See for example Bagnoli and Bergstrom (2005)

$(0,1)$ as shown in the previous lemma. It follows that $\mathrm{P}(\text{draw j}|\text{lie})\mathcal{M}'(\hat{t}_j(\varphi))\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} < \mathrm{P}(\text{draw j}|\text{lie})$ for all $j \leq k^*$. Therefore the sum

$$\sum_{j=1}^{k^*} \mathrm{P}(\text{draw j}|\text{lie})\mathcal{M}'(\hat{t}_j(\varphi))\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} < \sum_{j=1}^{k^*} \mathrm{P}(\text{draw j}|\text{lie}) = 1.$$

We now show that the second term in equation (13) is weakly smaller than zero. First note that $\sum_{j\leq k^*} \frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi} = 0$, as the probabilities always add up to one. In the case where $k^* = 1$ it then straightforwardly follows that the second term is zero. For cases $k^* > 1$, denote by S^+ and S^- the set of all states $j \leq k^*$ for which with $\frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi} > 0$ and $\frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi} \leq 0$ respectively. We know that $\sum_{j\in S^+} \frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi} = -\sum_{j\in S^-} \frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi}$. Therefore,

$$\begin{aligned}
&\sum_{j=1}^{k^*} \frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi}\mathcal{M}(\hat{t}_j(\varphi)) < 0 \text{ iff} \\
&\sum_{j\in S^+} \frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi}\mathcal{M}(\hat{t}_j(\varphi)) < -\sum_{j\in S^-} \frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi}\mathcal{M}(\hat{t}_j(\varphi)),
\end{aligned}$$

which in particular holds if $\min_{j\in S^-}\{\mathcal{M}(\hat{t}_j(\varphi))\} > \max_{j\in S^+}\{\mathcal{M}(\hat{t}_j(\varphi))\}$. We know that $\mathcal{M}(\hat{t}_1(\varphi)) > ... > \mathcal{M}(\hat{t}_{k^*}(\varphi))$. Hence, it is sufficient to show that there exists a $z \in \{1, k^*\}$ so that $S^- = \{j | j \in \mathcal{K}, j \leq z\}$. In words, we need to show that a marginal increase in φ leads to a relative decline in the proportion of liars who draw low states, as compared to liars who draw high states. The derivative $\frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi}$ has the same sign as

$$\frac{\frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi}}{\mathrm{P}(\text{draw j}|\text{lie})} = \frac{f(\hat{t}_j)}{F(\hat{t}_j)}\frac{\partial \hat{t}_j}{\partial \varphi} - c,$$

where $c = \sum_{l\leq k^*} f(\hat{t}_l)\frac{\partial \hat{t}_l}{\partial \varphi}$ is a constant which is the same for all $j \leq k^*$. This implies that $\frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi} < 0$ if and only if

$$\frac{f(\hat{t}_j(\varphi))}{F(\hat{t}_j(\varphi))}\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} < c.$$

It follows that if

$$\frac{f(\hat{t}_j(\varphi))}{F(\hat{t}_j(\varphi))}\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} < \frac{f(\hat{t}_k(\varphi))}{F(\hat{t}_k(\varphi))}\frac{\mathrm{d}\hat{t}_k}{\mathrm{d}\varphi} \text{ for all } j < k < k^*, \tag{14}$$

then there is a z so that $S^- = \{j | j \in \mathbb{N}_+, j \leq z\}$. Inequality (14) holds, because by log-concavity $\frac{f(\hat{t}_j(\varphi))}{F(\hat{t}_j(\varphi))} < \frac{f(\hat{t}_k(\varphi))}{F(\hat{t}_k(\varphi))}$ and by the previous lemma, $\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} < \frac{\mathrm{d}\hat{t}_k}{\mathrm{d}\varphi}$. Thus, if there exists a z with $\frac{\partial \mathrm{P}(\text{draw z}|\text{lie})}{\partial \varphi} \leq 0$, then $\frac{\mathrm{dP}(\text{draw j}|\text{lie})}{\mathrm{d}\varphi} < 0$ for all $j < z$. Existence of z follows from the fact that the derivatives always have to add up to zero. The second term in the derivative is therefore weakly negative. This concludes the proof of the claim that $\mathcal{L}'(\varphi) < 1$.

(iii) The assumptions on the payoff function ensure that some agents always lie regardless of the reputational penalty and that some agents always tell the truth even if there is no reputational penalty from lying. The first implies that $\mathcal{L}(0) > 0$, as some types will lie even if $\varphi = 0$. The second property implies that $\mathcal{L}(E(t)) < E(t)$ as the types with the highest lying cost tell the truth even if $\varphi = E(t)$. Since $\mathcal{L}(\varphi)$ is continuous and $\mathcal{L}'(\varphi) \in (0,1)$, it then follows that there exists a unique fixed point $\xi \in (0, E(t))$ such that $\mathcal{L}(\xi) = \xi$. ■

Proof of proposition 1. I omit the proofs for claims $(i) - (iv)$ in the proposition as they have been

given in the text and instead focus on the existence and uniqueness of equilibrium.

Claim 1: For every φ there exists a unique threshold value k^* which is the maximum integer $j \in \{1,...,K-1\}$ such that $y(K)+\mu\varphi \geq y(j)+\mu E(t)$. Assume that there is a k^* for which $y(K)+\mu\varphi \leq y(k^*)+\mu E(t)$. But then, individuals can profitably deviate and report k^*, as in such an equilibrium $\mathcal{R}_{k^*} > E(t) > \varphi$, and hence

$$y(k^*)+\mu\mathcal{R}_{k^*} > y(K)+\mu\varphi.$$

This establishes that for any k^*, $y(K)+\mu\varphi > y(k^*)+\mu E(t)$.

To see that k^* is the largest integer, consider a case where $k^* < k'$ and

$$y(K)+\mu\varphi > y(k')+\mu E(t).$$

Since k' is now being lied at, $\mathcal{R}_{k'} < E(t)$. The inequality is a contradiction to the condition that in any such equilibrium, $y(K)+\mu\varphi = y(k')+\mu\mathcal{R}_{k'}$.

Claim 2: For every $\varphi \in (\xi, E(t))$, the fraction of agents who lie is defined by a function $S(\varphi) = \frac{1}{K}\sum_{i=1}^{k^*} F(\hat{t}_j(\varphi))$. S is continuous with $S'(\varphi) > 0$. The first part follows because agents only lie if they draw a state smaller or equal k^* and lie if they have a lying cost lower than the threshold cost $\hat{t}_j(\varphi)$. Therefore, the fraction of agents who are liars is given by S
Continuity of S follows from very similar arguments as the ones that were used to show the continuity of $\mathcal{L}(\varphi)$. Consider a φ' such that $y(K)+\mu\varphi' = y(k')+\mu E(t)$ and subtract a minimal amount $\varepsilon > 0$ to $\varphi' - \varepsilon$. The threshold state now becomes $k'-1$, and

$$\begin{aligned} S(\varphi'-\varepsilon) &= \frac{1}{K}\sum_{j=1}^{k'-1} F(\hat{t}_j(\varphi'-\varepsilon)) \\ &= \frac{1}{K}\sum_{j=1}^{k'} F(\hat{t}_j(\varphi'-\varepsilon)) - \frac{1}{K}F(\hat{t}_{k'}(\varphi')). \end{aligned}$$

Now, since $\hat{t}_j(\varphi)$ and $F(t)$ are continuous and $\lim_{\varepsilon\to 0}\frac{1}{K}F(\hat{t}_{k'}(\varphi'-\varepsilon)) = F(0) = 0$, it follows that $\lim_{\varepsilon\to 0} S(\varphi'-\varepsilon) = S(\varphi')$. Moreover, $F'(t) > 0$ and $\hat{t}'_j(\varphi) > 0$ for $j \leq k^*$ and thus $S'(\varphi) > 0$ – the supply of lies increases in the reputation of the highest state.

Claim 3: For every $\varphi \in (\xi, E(t))$, $D(\varphi) = \frac{1}{K}\sum_{j=k^*+1}^{K}\frac{1-r_j(\varphi)}{r_j(\varphi)}$ is continuous with $D'(\varphi) < 0$. In equilibrium, $D(\varphi) = \mathrm{P}(lie)$. The fraction of liars that report a state larger than k^* is

$$\sum_{j=k^*+1}^{K} \mathrm{P}(\text{report } j) \times \mathrm{P}(\text{lie}|\text{report } j). \tag{15}$$

We defined $r_j = P(\text{truth}|\text{report } j)$. By Bayes' Rule,

$$r_j = \frac{\mathrm{P}(\text{report } j \wedge \text{truth})}{\mathrm{P}(\text{report } j)} \text{ for } j > k^*.$$

Observe that in equilibrium exactly $\frac{1}{K}$ agents report each state $j > k^*$ truthfully. Thus, we can rearrange the above equation to

$$\mathrm{P}(\text{report } j) = \frac{1}{K}\frac{1}{r_j}.$$

Plugging into (15), we arrive at the following expression

$$\sum_{j=k^*+1}^{K} \text{P(report } j) \times \text{P(lie|report } j) = \frac{1}{K} \sum_{j=k^*+1}^{K} \frac{1-r_j}{r_j}.$$

We can derive an expression for r_j depending on φ by noting that,

$$E(t|j) = r_j E(t) + (1-r_j)\mathcal{L}(\varphi) \text{ for all } j > k^*$$

and use the indifference conditions to replace $E(t|j) = \varphi + \frac{\Delta(K,j)}{\mu}$ for $j > k^*$ to derive

$$r_j(\varphi) = \frac{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\varphi)}{E(t) - \mathcal{L}(\varphi)}.$$

Finally, we define

$$D(\varphi) \equiv \frac{1}{K} \sum_{j=k^*+1}^{K} \frac{1-r_j(\varphi)}{r_j(\varphi)} = \frac{1}{K} \sum_{j=k^*+1}^{K} \frac{E(t) - \varphi + \Delta(K,j)/\mu}{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\varphi)}.$$

To see that $D(\varphi)$ is continuous, we again have to consider a value φ' such that $y(K) + \mu\varphi' = y(k') + \mu E(t)$. The threshold state for φ' is k', while it is $k'-1$ for $\varphi - \varepsilon$, where ε is an arbitrarily small but positive scalar. After plugging in the equilibrium conditions we have

$$\begin{aligned} D(\varphi' - \varepsilon) &= \frac{1}{K} \sum_{j=k^*}^{K} \frac{E(t) - (\varphi' - \varepsilon + \frac{\Delta(K,j)}{\mu})}{\varphi' - \varepsilon + \frac{\Delta(K,j)}{\mu} - \mathcal{L}(\varphi' - \varepsilon)} \\ &= \frac{1}{K} \sum_{j=k^*+1}^{K} \frac{E(t) - (\varphi' - \varepsilon + \frac{\Delta(K,j)}{\mu})}{\varphi' - \varepsilon + \frac{\Delta(K,j)}{\mu} - \mathcal{L}(\varphi' - \varepsilon)} + \frac{1}{K} \frac{E(t) - (\varphi' - \varepsilon + \frac{\Delta(K,k')}{\mu})}{\varphi' - \varepsilon + \frac{\Delta(K,k')}{\mu} - \mathcal{L}(\varphi' - \varepsilon)}. \end{aligned}$$

In the limit where $\varepsilon \to 0$, $\varphi' - \varepsilon + \frac{\Delta(K,k')}{\mu} = E(t)$. The second term in the equation above therefore becomes zero in the limit zero, and as $\mathcal{L}(\varphi)$ and $E(t|j)$ are continuous, $D(\varphi' - \varepsilon) \to D(\varphi')$. $D(\varphi)$ is decreasing in φ. Taking the derivative,

$$D'(\varphi) = \frac{1}{K} \sum_{j=k^*+1}^{K} \frac{\mathcal{L}'(\varphi)(E(t) - E(t|j)) - (E(t) - \mathcal{L}(\varphi))}{(E(t|j) - \mathcal{L}(\varphi))^2}.$$

A sufficient condition for the derivative to be negative is that the numerator in the fraction above is negative for every $j > k^*$. Since $E(t|j) > \mathcal{L}(\varphi)$, a sufficient condition for the derivative to be negative is that

$$\mathcal{L}'(\varphi) < 1,$$

which was shown in lemma 2.

Claim 4: There exists a unique $\varphi^* \in (\zeta, E(t))$ such that $D(\varphi^*) = S(\varphi^*)$. From the previous claims, it follows that $D(\varphi)$ and $S(\varphi)$ are both continuous functions with $D'(\varphi) < 0$ and $S'(\varphi) > 0$. The intermediate value theorem guarantees a unique φ^* such that $D(\varphi^*) = S(\varphi^*)$. For existence of φ^*, observe that the parameter assumptions guarantee that $S(\varphi) \in (0,1)$ for all $\varphi \in (\xi, E(t))$. When $\varphi \to \xi$, $\varphi - \mathcal{L}(\varphi) \to 0$ and thus $\lim_{\varphi \to \xi} D(\varphi) = \infty$. In the case where $\varphi \to E(t)$, $k^* = K-1$ and thus

$$\lim_{\varphi \to E(t)} D(\varphi) = \lim_{\varphi \to E(t)} \frac{1}{K} \frac{E(t) - \varphi}{\varphi - \mathcal{L}(\varphi)} = 0.$$

It follows that

$$\lim_{\varphi\to\xi}[D(\varphi)-S(\varphi)]>0, \text{ and } \lim_{\varphi\to E(t)}[D(\varphi)-S(\varphi)]<0.$$

As the difference is continuous and strictly decreasing there exists a unique $\varphi^*\in(\zeta,E(t))$ such that $D(\varphi^*)=S(\varphi^*)$. ■

A.2 Proof of proposition 2

Claim 1: $K\in\mathcal{Q}_l$ Assume the opposite and observe that K is always reported dishonestly by some agent. It follows that

$$y(K)+\mu_h R_K\geq y(j)+\mu_h\mathcal{R}_j$$

and

$$y(j)+\mu_l R_j>y(K)+\mu_l R_K \text{ for some } j\neq K.$$

Combining both inequalities yields $\mu_l>\frac{\Delta(K,j)}{\mathcal{R}_j-\mathcal{R}_K}\geq\mu_h$, a contradiction to $\mu_l<\mu_h$.

Claim 2: The intersection $\mathcal{Q}_l\cap\mathcal{Q}_h$ is at most a singleton. Assume the opposite. Then, there exist at least two states $j,k\in\mathcal{Q}_l,\mathcal{Q}_h$. For both

$$\begin{aligned}y(k)+\mu_h R_k&=y(j)+\mu_h R_j,\\ y(k)+\mu_l R_k&=y(j)+\mu_l R_j.\end{aligned}$$

Combining both equalities implies $\mu_h=\frac{\Delta(k,j)}{\mathcal{R}_j-\mathcal{R}_k}=\mu_l$, which is a contradiction to $\mu_l<\mu_h$.

Claim 3: $\mathcal{Q}_l$ *and* $\mathcal{Q}_h$ *can intersect only at the minimum* $j\in\mathcal{Q}_l$. For $|\mathcal{Q}_h|=1$ this is trivial. For $|\mathcal{Q}_h|>1$, we know from claim 1 that

$$y(K)+\mu_l R_K=y(j)+\mu_l\mathcal{R}_j \text{ for all } j\in\mathcal{Q}_l.$$

Since $\mu_h>\mu_l$, this implies that for h-types,

$$y(K)+\mu_h\mathcal{R}_K<y(K-1)+\mu_h\mathcal{R}_{K-1}<\ldots<y(m)+\mu_h\mathcal{R}_m,$$

where m is the minimum element in $\mathcal{Q}_l$. Therefore, any h-liar prefers to report m.

A.3 Proof of proposition 3

We know from propositon 1 that φ^* is implicitly defined by $D(\varphi^*,p,q)-S(\varphi^*,p,q)=0$. In the restricted game, the demand and supply functions are the following:

$$\begin{aligned}D(\varphi,p,q)&=p\times\frac{E(t)-\varphi}{\varphi-\mathcal{L}(\varphi,q)}\\ S(\varphi,p,q)&=(1-p)q\times F(\hat{t}_G(\varphi))+(1-p)(1-q)\times F(\hat{t}_B(\varphi)).\end{aligned}\tag{16}$$

Implicitly differentiating the equilibrium condition gives

$$\frac{\mathrm{d}\varphi^*}{\mathrm{d}p}=\frac{D_p(\varphi^*,p,q)-S_p(\varphi^*,p,q)}{S_\varphi(\varphi^*,p,q)-D_\varphi(\varphi^*,p,q)}.$$

Also from proposition 1 we know that denominator is positive. Taking derivatives of the functions specified in (16) gives (note that $\mathcal{L}(\varphi,q)$ only depends on q, but not on p)

$$D_p(\varphi,p,q) = \frac{E(t)-\varphi}{\varphi-\mathcal{L}(\varphi,q)} > 0$$
$$S_p(\varphi,p,q) = -qF(\hat{t}_G(\varphi)) - (1-q)F(\hat{t}_B(\varphi)) < 0.$$

We conclude that $\frac{\mathrm{d}\varphi^*}{\mathrm{d}p} > 0$.

A.4 Proof of proposition 4

(i) From equations (16), it follows that

$$D_q(\varphi,p,q) = p \times \frac{\mathcal{L}_q(\varphi,q)}{\varphi-\mathcal{L}(\varphi,q)} D(\varphi,p,q),$$
$$S_q(\varphi,p,q) = (1-p)\left[F(\hat{t}_G(\varphi)) - F(\hat{t}_B(\varphi))\right].$$

Since $\hat{t}_G(\varphi) > \hat{t}_B(\varphi)$ and $F(\hat{t})$ is increasing, it follows that $S_q(\varphi,p,q) > 0$. The sign of $D_q(\varphi,p,q)$ depends on

$$\mathcal{L}_q(\varphi,q) = \frac{\partial \mathrm{P}(\text{draw } \textit{Blue}|\text{lie})}{\partial q}\left[\mathcal{M}(\hat{t}_B(\varphi)) - \mathcal{M}(\hat{t}_G(\varphi))\right],$$

where $\mathrm{P}(\text{draw } \textit{Blue}|\text{lie}) = \frac{(1-q)F(\hat{t}_B(\varphi))}{(1-q)F(\hat{t}_B(\varphi))+qF(\hat{t}_G(\varphi))}$, which is decreasing in q. Therefore $D_q(\varphi,p,q)$ is also positive. Since both $S_q(\varphi,p,q)$ and $D_q(\varphi,p,q)$ are positive it directly follows that $\mathrm{P}(\text{lie})$ is higher the higher q.

(ii) In equilibrium

$$\varphi^* = \underbrace{\frac{pE(t) + (1-p)\left[(1-q)F(\hat{t}_B(\varphi^*))\mathcal{M}(\hat{t}_B(\varphi^*)) + qF(\hat{t}_G(\varphi^*))\mathcal{M}(\hat{t}_G(\varphi^*))\right]}{p + (1-p)\left[(1-q)F(\hat{t}_B(\varphi^*)) + qF(\hat{t}_G(\varphi^*))\right]}}_{g(\varphi^*,p,q)}.$$

Define

$$\Phi(\varphi,p,q) = \varphi - g(\varphi,p,q).$$

We can implicitly differentiate $\Phi(\varphi^*,p,q)$ to get

$$\frac{\mathrm{d}\varphi^*}{\mathrm{d}q} = \frac{g_q(\varphi^*,p,q)}{1-g_\varphi(\varphi^*,p,q)}.$$

The denominator is always positive as $g_\varphi(\varphi,p,q) < 1$. The sign of the whole derivative is thus determined by the sign of the derivative in the numerator, which becomes

$$g_q(\varphi,p,q) = \frac{p\left[F(\hat{t}_B(\varphi))(\mathcal{M}(\hat{t}_B(\varphi)) - E(t)) - F(\hat{t}_G(\varphi))(\mathcal{M}(\hat{t}_G(\varphi)) - E(t))\right] + (1-p)F(\hat{t}_B(\varphi))F(\hat{t}_G(\varphi))(\mathcal{M}(\hat{t}_B(\varphi)) - \mathcal{M}(\hat{t}_G(\varphi)))}{\left[p + (1-p)\left[(1-q)F(\hat{t}_B(\varphi^*)) + qF(\hat{t}_G(\varphi^*))\right]\right]^2}.$$

The second term in the numerator is always negative, while the first term is negative if

$$\begin{aligned} F(\hat{t}_G)\left[E(t)-\mathcal{M}(\hat{t}_G(\varphi))\right] &< F(\hat{t}_B(\varphi))\left[E(t)-\mathcal{M}(\hat{t}_B(\varphi))\right] \\ \Rightarrow E(t)\left[F(\hat{t}_G(\varphi))-F(\hat{t}_B(\varphi))\right] &< \int_0^{\hat{t}_G(\varphi)} sf(s)\,\mathrm{d}s - \int_0^{\hat{t}_B(\varphi)} sf(s)\,\mathrm{d}s = \int_{\hat{t}_B(\varphi)}^{\hat{t}_G(\varphi)} sf(s)\,\mathrm{d}s \\ \Rightarrow E(t) &< E(t|t\in(\hat{t}_B(\varphi),\hat{t}_G(\varphi))). \end{aligned}$$

A sufficient condition for the last inequality to hold is that $\hat{t}_B(\varphi^*) > E(t)$.

B Example of a non-symmetric equilibrium

This section provides an example of an equilibrium where the reports of liars depends on their lying cost. Consider a setup with $K = 3$ and the following strategy profile:

$$\begin{aligned} &s(j|j,t) = 1 \text{ if } j > 1, \\ &s(3|1,t) = 1 \text{ if } t \le \hat{t}_a \\ &s(2|1,t) = 1 \text{ if } t \in (\hat{t}_a,\hat{t}_b] \\ &s(1|1,t) = 1 \text{ if } t \ge \hat{t}_b. \end{aligned}$$

That is, agents lie only if they draw 1. There are two quality segments of liars. Liars with the worst quality report the highest state and other liars report the middle state.

Assume preferences are uniformly distributed between zero and $T > 0$. The equilibrium reputations are

$$\begin{aligned} \mathcal{R}_1(T,\hat{t}_b) &= \frac{T+\hat{t}_b}{2} \\ \mathcal{R}_2(T,\hat{t}_a,\hat{t}_b) &= \frac{1}{2}\times\frac{T^2+(\hat{t}_a+\hat{t}_b)(\hat{t}_b-\hat{t}_a)}{T+\hat{t}_b-\hat{t}_a} \\ \mathcal{R}_3(T,\hat{t}_a) &= \frac{1}{2}\times\frac{T^2+\hat{t}_a^2}{T+\hat{t}_a}. \end{aligned}$$

The equilibrium is characterized by two threshold values $(\hat{t}_a,\hat{t}_b)$ and two indifference conditions. The first is that the agent of type $(1,\hat{t}_b)$ must be indifferent between lying and truthtelling;

$$\begin{aligned} y(1)+\mu\mathcal{R}_1(T,\hat{t}_b) &= y(3)+\mu\mathcal{R}_3(t,\hat{t}_a)-\hat{t}_b \\ \Rightarrow \hat{t}_b &= \frac{1}{1+\mu/2}\left(\Delta(3,1)+\mu(R_3(T,\hat{t}_a)-T)\right). \end{aligned} \tag{17}$$

The second equilibrium condition is that liars must be indifferent between reporting states 2 and 3;

$$y(3)+\mu\mathcal{R}_3(T,\hat{t}_a) = y(2)+\mu\mathcal{R}_2(T,\hat{t}_a,\hat{t}_b). \tag{18}$$

Consider parameter values $\mu = 1$, $T = 7$, $y(1) = 0$, $y(2) = 4.9$, and $y(3) = 5$. We can plug equation (17) into equation (18) and solve for $\hat{t}_a$. The resulting parameter values are $\hat{t}_a \approx 1.12$ and $\hat{t}_b \approx 3.06$, which imply that each of the states are reported (from low to high) with frequencies 18.75%, 42.57%, and 38.68%. Note that in this example, the second-highest state is reported with a higher frequency than the highest state. In the symmetric equilibrium with homogeneous image concerns the reporting frequencies are monotonely increasing in j. Therefore, the example induces a different reporting frequency than the one induced by symmetric equilibrium.

C Details on the maximum likelihood estimation

The individual likelihood contributions are derived at as follows: Consider a participant who reports *Green* and holds some belief b_i. Their probability of truth-telling, given the model parameters, is $\mathrm{P}(\text{report } \textit{Green} | \text{draw } \textit{Green}, b_i, t_i)$ and the probability of observing such a belief is $g(b_i - \beta_{\text{treatment}} - \rho t_i)$, where g is the normal p.d.f. (c.d.f.) if b_i is between zero and one (equal to zero or one). The total likelihood contribution of this participant is then derived by integrating out the unobserved lying cost t_i, whose c.d.f. is denoted by $F(t_i)$;

$$\int \mathrm{P}(\text{report } \textit{Green} | \text{draw } \textit{Green}, b_i, t_i) g(b_i - \beta_{\text{treatment}} - \rho t_i) \, \mathrm{d}F(t_i).$$

The likelihood contribution of an individual who reported *Red* is more ambiguous, as they could have either reported truthfully or lied. Thus, any participant who reports red could have (i) drawn *Red* and told the truth, (ii) drawn *Blue* and lied, or (iii) drawn *Green* and lied. This implies that the likelihood contribution of an individual who reported *Red* becomes

$$\begin{aligned} &\int \big(p + (1-p)\big[q\mathrm{P}(\text{report } \textit{Red} | \text{draw } \textit{Green}, b_i, t_i) + (1-q)\mathrm{P}(\text{report } \textit{Red} | \text{draw } \textit{Blue}, b_i, t_i)\big]\big) \\ &\times g(b_i - \beta_{\text{treatment}} + \rho t_i) \, \mathrm{d}F(t_i). \end{aligned}$$

We can sometimes further exploit session-level variation on the colors that liars drew to increase the efficiency of the estimator. For example, if in one session all individuals who drew *Blue* told the truth, we can set q to zero (keeping the probability of having drawn *Green* at $(1-p)(1-q)$) for observations from that session only.

Taking logs and summing up the individual likelihood contributions gives the log-likelihood function, which is then maximized with respect to the unknown parameters.

D Experimental instructions

Instructions were translated from German.

D.1 BASE

D.1.1 Handout 1

Instructions

12 people will take part in today's experiment. The experiment consists of a lottery. You will first draw a prize in secret which you will have to register. There are three different lottery prizes, which we will call *Red*, *Blue*, and *Green*. You will receive an amount of money depending on the color you report. At the end of the experiment, you will tell all other participants and the experimenter which prize you won.
The table below displays all possible payouts which you can earn in the lottery.

Reported color	Red	Blue	Green
Payout	12€	10€	2€

The lottery will be the only task in today's experiment. It will proceed as follows: We will come and ask you to draw an envelope from a box. There are two winning codes in the envelope. Every winning code consists of a digit code. You have to report the code to receive your prize from the experimenter.
Nine participants will draw an envelope which contains the winning codes for *Red* and *Blue*. The remaining three participants will draw an envelope with the winning codes for *Red* and *Green*.
There exists one winning code for each color. That is, for example, all participants see exactly the same winning code for *Red*.
Every envelope additionally contains a lot, which can either be a "win" or a "loss".
If "win" is written on your lot: Denote the winning code for *Red* on the scrap sheet in front of you.
If "loss" is written on your lot: Denote the winning code for the other color on the scrap sheet in front of you.
There are four winning lots in total, three of which are in the *Red-Blue* envelopes and one which is in the *Red-Green* envelopes. It follows that four participants will be asked to report *Red*, six will be asked to report *Blue*, and two will be asked to report *Green*.
Figure 1 (on the extra sheet of paper on your desk) presents the lottery graphically: There are twelve envelopes of which you randomly draw one. Every envelope contains two winning codes. There are nine envelopes that contain winning codes for *Red* and *Blue*. Six of these nine envelopes contain a loss (L) and three contain a win (W). The remaining three envelopes contain winning codes for *Red* and *Green*. There are two losses and one win contained in these envelopes.
Please note the following properties of the lottery:

- 4 participants win *Red*, 6 participants win *Blue* and 2 participants win *Green*.
- Every person sees 2 out of 3 possible winning codes.
- All participants see the winning code for *Red*.

- No participant who wins *Green* sees the winning code of *Blue*.
- No participant who wins *Blue* sees the winning code of *Green*.

After all participants denoted their code, we will come with a sealed box into which you can throw your envelope and lot. We will destroy all envelopes and lots after the experiment. This procedure ensures that neither the remaining participants, nor the experimenters will find out which envelope you drew.

Report your winning code thereafter to the computer program. The program will register your color and you will receive the corresponding euro amount at the end of the experiment.

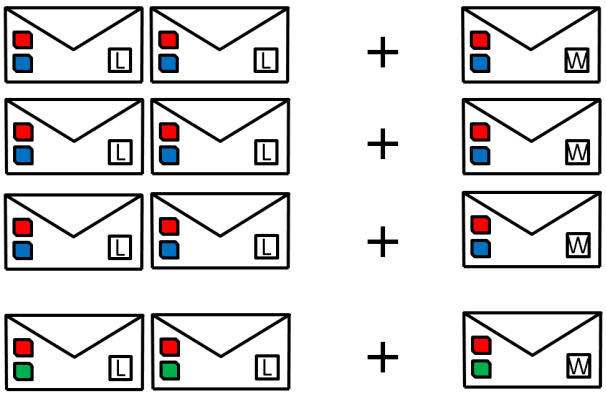
Before you leave the room to pick up your payout, you will see your payout and reported color on your screen. We will then ask you to stand up. Thereafter, we will come one after another to every participant and will ask you to read out loudly the color reported by you.

Take your time to read through the instructions again and make sure that you understood everything. Please raise your hand if you have questions and we will come to you.

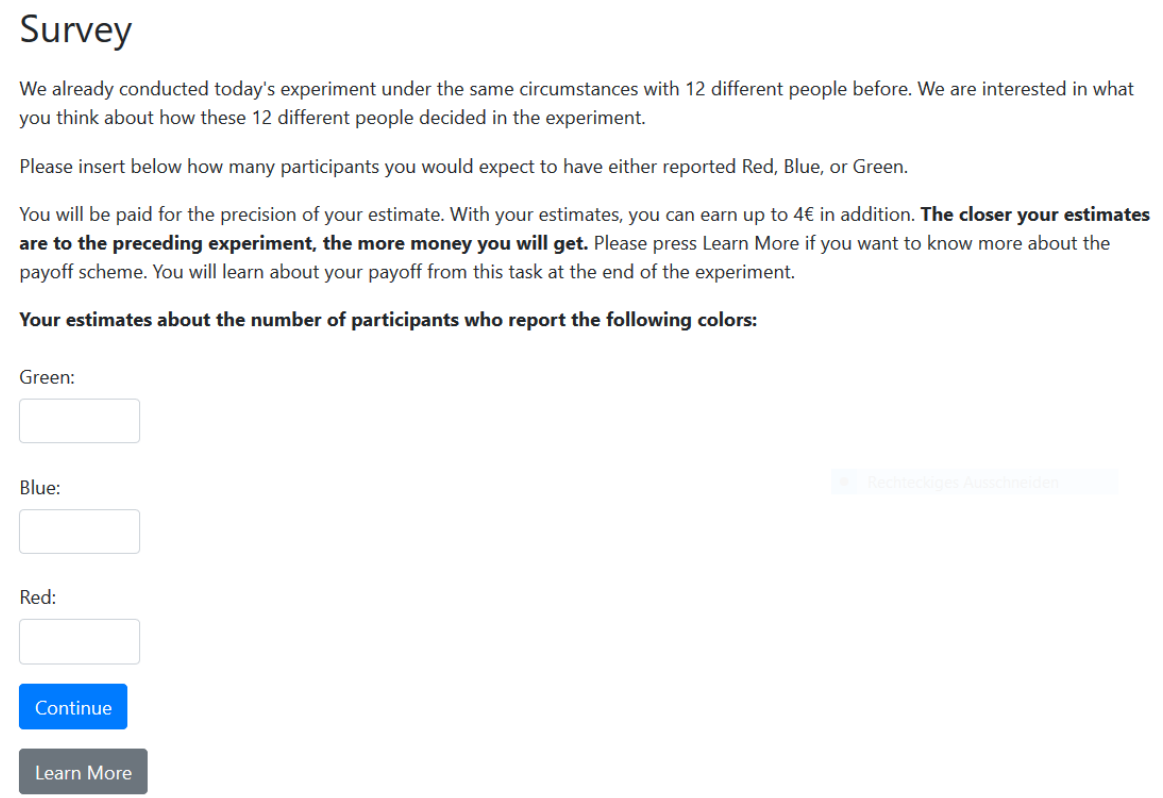
Please press proceed on your screen once you are finished. Your will then answer some control questions before the lottery begins.

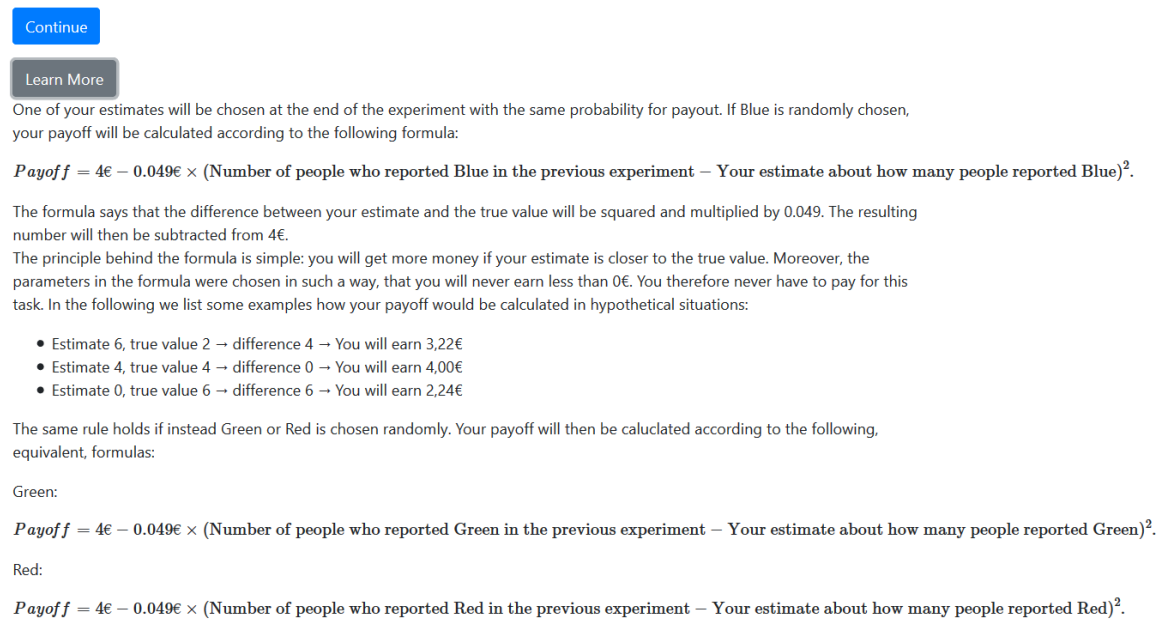
D.1.2 Handout 2

Figure 1



D.1.3 Screenshots of belief elicitation





D.2 HQ

D.2.1 Handout 1

Instructions

12 people will take part in today's experiment. The experiment consists of a lottery. You will first draw a prize in secret which you will have to register. There are three different lottery prizes, which we will call *Red*, *Blue*, and *Green*. You will receive an amount of money depending on the color you report. At the end of the experiment, you will tell all other participants and the experimenter which prize you won.
The table below displays all possible payouts which you can earn in the lottery.

Reported color	Red	Blue	Green
Payout	12€	10€	2€

The lottery will be the only task in today's experiment. It will proceed as follows: We will come and ask you to draw an envelope from a box. There are two winning codes in the envelope. Every winning code consists of a digit code. You have to report the code to receive your prize from the experimenter.
Three participants will draw an envelope which contains the winning codes for *Red* and *Blue*. The remaining nine participants will draw an envelope with the winning codes for *Red* and *Green*.
There exists one winning code for each color. That is, for example, all participants see exactly the same winning code for *Red*.
Every envelope additionally contains a lot, which can either be a "win" or a "loss".
If "win" is written on your lot: Denote the winning code for *Red* on the scrap sheet in front of you.
If "loss" is written on your lot: Denote the winning code for the other color on the scrap sheet in front of you.
There are four winning lots in total, one of which are in the *Red-Blue* envelopes and three which are in the *Red-Green* envelopes. It follows that four participants will be asked to report *Red*, two will be asked to report *Blue*, and six will be asked to report *Green*.
Figure 1 (on the extra sheet of paper on your desk) presents the lottery graphically: There are twelve envelopes of which you randomly draw one. Every envelope contains two winning codes. There are three envelopes that contain winning codes for *Red* and *Blue*. Two of these three envelopes contain a loss (L) and one contains a win (W). The remaining nine envelopes contain winning codes for *Red* and *Green*. There are six losses and three wins contained in these envelopes.
Please note the following properties of the lottery:

- 4 participants win *Red*, 2 participants win *Blue* and 6 participants win *Green*.
- Every person sees 2 out of 3 possible winning codes.
- All participants see the winning code for *Red*.
- No participant who wins *Green* sees the winning code of *Blue*.
- No participant who wins *Blue* sees the winning code of *Green*.

After all participants denoted their code, we will come with a sealed box into which you can throw your envelope and lot. We will destroy all envelopes and lots after the experiment. This procedure ensures that neither the remaining participants, nor the experimenters will find out which envelope you drew.

Report your winning code thereafter to the computer program. The program will register your color and you will receive the corresponding euro amount at the end of the experiment.

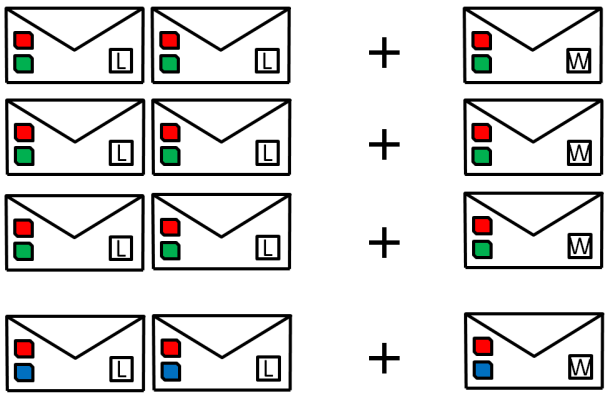
Before you leave the room to pick up your payout, you will see your payout and reported color on your screen. We will then ask you to stand up. Thereafter, we will come one after another to every participant and will ask you to read out loudly the color reported by you.

Take your time to read through the instructions again and make sure that you understood everything. Please raise your hand if you have questions and we will come to you.

Please press proceed on your screen once you are finished. Your will then answer some control questions before the lottery begins.

D.2.2 Handout 2

Figure 3



D.3 LP

D.3.1 Handout 1

Instructions

12 people will take part in today's experiment. The experiment consists of a lottery. You will first draw a prize in secret which you will have to register. There are three different lottery prizes, which we will call *Red*, *Blue*, and *Green*. You will receive an amount of money depending on the color you report. At the end of the experiment, you will tell all other participants and the experimenter which prize you won.

The table below displays all possible payouts which you can earn in the lottery.

Reported color	Red	Blue	Green
Payout	12€	10€	2€

The lottery will be the only task in today's experiment. It will proceed as follows: We will come and ask you to draw an envelope from a box. There are two winning codes in the envelope. Every winning code consists of a digit code. You have to report the code to receive your prize from the experimenter.

Nine participants will draw an envelope which contains the winning codes for *Red* and *Blue*. The remaining three participants will draw an envelope with the winning codes for *Red* and *Green*.

There exists one winning code for each color. That is, for example, all participants see exactly the same winning code for *Red*.

Every envelope additionally contains a lot, which can either be a "win" or a "loss".

If "win" is written on your lot: Denote the winning code for *Red* on the scrap sheet in front of you.

If "loss" is written on your lot: Denote the winning code for the other color on the scrap sheet in front of you.

There is one winning lot in total, which either is in a *Red-Blue* envelope or in a *Red-Green* envelope. It follows that one participant will be asked to report *Red*, eight or nine participants will be asked to report *Blue*, and two or three participants will be asked to report *Green*.

Figure 1 (on the extra sheet of paper on your desk) presents the lottery graphically: There are twelve envelopes of which you randomly draw one. Every envelope contains two winning codes. There are nine envelopes that contain winning codes for *Red* and *Blue*. At least eight of these nine envelopes contain a loss (L) and at most one contains a win (W). The remaining three envelopes contain winning codes for *Red* and *Green*. There are at least two losses and at most one win contained in these envelopes.

Note that there is exactly one winning lot among the twelve envelopes. That means that if the win is contained in a *Red-Blue* envelope, all *Red-Green* envelopes will contain a loss. Conversely it also holds that, if the win is contained in a *Red-Green* envelope, that all *Red-Blue* envelopes contain losses. The win is contained with equal probability in one of the twelve envelopes.

Please note the following properties of the lottery:

- 1 participant wins *Red*, 8 to 9 participants win *Blue* and 2 to 3 participants win *Green*.
- Every person sees 2 out of 3 possible winning codes.
- All participants see the winning code for *Red*.
- No participant who wins *Green* sees the winning code of *Blue*.
- No participant who wins *Blue* sees the winning code of *Green*.

After all participants denoted their code, we will come with a sealed box into which you can throw your envelope and lot. We will destroy all envelopes and lots after the experiment. This procedure ensures that neither the remaining participants, nor the experimenters will find out which envelope you drew.

Report your winning code thereafter to the computer program. The program will register your color and you will receive the corresponding euro amount at the end of the experiment.

Before you leave the room to pick up your payout, you will see your payout and reported color on your screen. We will then ask you to stand up. Thereafter, we will come one after another to every participant and will ask you to read out loudly the color reported by you.

Take your time to read through the instructions again and make sure that you understood everything. Please raise your hand if you have questions and we will come to you.

Please press proceed on your screen once you are finished. Your will then answer some control questions before the lottery begins.

D.3.2 Handout 2

Figure 1