Deterrence Effects of Enforcement Schemes: an Experimental Study

Marina Agranov*and Anastasia Buyalskaya†

February 12, 2021

Abstract

Private and public organizations are interested in finding effective ways to reduce crime and promote ethical behavior without investing heavy resources into monitoring and compliance. In this paper, we study experimentally how revealing different information about a fine distribution affects deterrence of an undesirable behavior. We use a novel, incentive-compatible elicitation method to observe subjects lying - the undesirable behavior - and quantify the extent to which this behavior responds to information structures. We find that punishment schemes which communicate only partial information - the minimum fine in particular - are more effective than full information schemes at deterring lying. We explore the mechanism driving this result and link it to subjects' beliefs about their own versus the average expected fine in treatments with partial information.

Acknowledgements: The authors benefited from helpful feedback from three anonymous referees, Devdeepta Bose, Colin Camerer, Federico Echenique, Lindsey Gailmard, Sera Linardi, Luciano Pomatto, Pietro Ortoleva, Stephanie Wang, as well as audience members at the LA Area Theory Workshop and the ESA conference in Dijon (2019) and SABE virtual conference attendees (2020).

*California Institute of Technology, e-mail: magranov@hss.caltech.edu.
†California Institute of Technology, e-mail: abuyalsk@caltech.edu.

1 Introduction

In this paper, we investigate the effectiveness of different punishment schemes in preventing an undesirable behavior ("a crime"), where a punishment scheme is made up of a monitoring probability and the expected fine if monitored. This is an important question for any type of institution that wishes to promote ethical and honest behavior among its constituents (e.g. employees, citizens) but may be constrained in its ability to monitor all individuals all the time. We consider settings in which fines must remain within a moderate range, such that the value of a transgression might be higher than the cost of punishment for some individuals (e.g. potentially paying a few hundred dollars for a speeding ticket or a carpool violation might be worth getting to a hospital quicker or to an important meeting on time). Therefore, conditional on a monitoring probability and a fine range being fixed, we ask whether revealing *different information* about the *same fine distribution* can result in some schemes being more effective at deterring undesirable behavior.

To motivate our research question, Figure 1 presents various ways in which fines for small traffic-related violations are typically presented. Some of these signs specify the exact fine amount (e.g. '$1000 Fine For Animal Abandonment'), while others mention the minimum fine (e.g. 'Red Light Violation $336 Minimum Fine'), and yet others are even more vague, asserting that fines will be double the regular ones without explicitly listing them (e.g. 'Double Fine Zone'). Abstracting away from any legal reasons behind these frame choices, we provide some of the first empirical evidence comparing the effectiveness of these different frames. More generally, we ask whether people react differently to partial information about a fine distribution as compared with full information about the same distribution, and explore the mechanism behind this difference in behavior.

To achieve this goal, we conduct a series of laboratory experiments in which subjects have the opportunity to lie ("commit the crime"), an action that rewards them with higher earnings if they are not caught. In the experiment, each subject is allocated one of five cards numbered from 1 to 5 at random and is asked to report the card number she receives. A subject is paid twice the number she reports. In treatments with monitoring and punishment, subjects' reports are monitored with commonly known probability of 20% and a subject faces a fine if she is caught lying. In line with our research goal, we use a variety of monitoring treatments in which we alter the information communicated about the fine structure while keeping the expected value of the fine constant. These treatments include a fixed fine, a random fine (equal chance of receiving a high or low fine), a minimum fine, and a maximum fine. The distribution of fines in the latter three treatments is held constant,

Figure 1: Examples of Signs Containing Fine Information in Los Angeles County (2019)

Notes: The most recent blank California MUTCD Sign Charts can be found at: https://dot.ca.gov/programs/traffic-operations/sign-charts

with an expected value equal to that of the fixed fine treatment.[1]

Our experimental results show that a "partial information" frame in which the minimum fine is communicated is the most effective at deterring lying as compared with other schemes. This result holds true both when looking at the aggregated behavior of subjects across all cards as well as individual reports for intermediate cards, for which the opportunity cost of lying is in the middle range.[2] In addition, when subjects are split up into types, the minimum treatment is again most effective at deterring the "worst offenders," or those most likely to commit a crime regardless of the opportunity cost.

To replicate this result and pin down why the minimum fine framing is most effective, we conduct a follow-up experiment. In the follow-up experiment, we focus on the two partial information schemes, the minimum fine and the maximum fine, and elicit subjects'

[1]Many small offenses, such as littering, tend to be associated with fines of significantly larger magnitudes than the potential benefits associated with violating them. These large fines are usually coupled with very small probabilities of being monitored by authorities. In contrast, in our experiments the probability of being monitored is quite substantial and fixed at 20% while the fines do not exceed the highest payoff one can earn in this task. This discrepancy is, however, mitigated by the substitution effect between the severity of punishments and monitoring probabilities, which has been documented extensively in the experimental literature (see Alm, Jackson, and McKee (1992) and Alm, Sanchez, and De Juan (1995) for tax compliance experiments, Friesen (2012), Feess, Schildberg-Hörisch, Schramm, and Wohlschlegel (2015), and DeAngelo and Charness (2012) for 'lab crime' experiments). Building upon results of this literature, we opted for fines that do not exceed benefits of lying, and, thus, eliminate the need to endow subjects with experimental currency for this task, thus, separating out the response to punishment schemes, which is the main focus of the paper, from potential confounding effects.

[2]This middle range is characterized by the opportunity cost of lying being not too high, where none of the punishment schemes successfully deter lying, nor too low, where all the punishment schemes are equally effective at preventing lying.

reports for different cards as well as their beliefs about the fines they would face if caught lying. Given that partial information schemes do not induce nor control subjects' beliefs about the fine distribution, these beliefs become a natural suspect for driving behavior in the two treatments. Specifically, we ask subjects to report (1) their belief about the *average fine* previous subjects participating in this treatment faced when caught lying, and (2) their belief about what their *own fine* would be if they are caught lying. We use these two questions to classify subjects based on the difference between two reported beliefs.[3]

We find several important results. First, about half of participants believe that their own fine would be higher than the average fine in the population. Second, subjects' behavior in the card task is strongly correlated with beliefs about their *own fine* but not with beliefs about the average fine. Third, the average and the median belief about own fines in the partial information treatments are not statistically different from average fines in the full information treatments, which means that this is not why partial information treatments are more effective at deterring lying. At the same time, the minimum fine treatment induces higher beliefs about one's own fine compared with the maximum treatment, which coupled with the second result shows why the minimum scheme outperforms the maximum one in terms of reducing lying behavior. Finally, lying behavior in the cards task is negatively correlated with the belief that own fine will be higher than the average, and with the spread of beliefs, which is the difference between own and average fines.

We conclude by noting that our results are applicable to a variety of settings, from managerial implications to individual decision-making environments and public economics more generally.[4] In all these settings, increasing resources required for monitoring is often too costly or operationally not possible, while shrouding the way information is presented when communicating a fine distribution, such as using a minimum fine, provides an alternative way to increase compliance.[5] This focus on how information is framed in order to increase compliance has been the subject of recent field research: Fishbane et al. (2020)

[3]Our preferred interpretation of the difference between own and average fines is that it relates to a subject's ambiguity attitude. Those who believe their own fine to be higher than the average tend to be ambiguity averse, while those who believe that their own fine will be lower than the average are ambiguity loving, with the remaining group being ambiguity neutral. We discuss this interpretation in detail in Section 5. However, we note that these belief measures are non-standard in the literature and might encompass other potential forces unrelated to ambiguity attitude such as the belief about being treated 'fairly' relative to others as well as potentially miscalibrated beliefs about own frequency of lying compared with others.

[4]For a survey of the tax compliance literature, see Slemrod (2019).

[5]This is perhaps one of several reasons why law enforcement officials use signs which advertise a "minimum" fine in lieu of an average fine. The other benefit of a minimum sign includes the fact that the sign requires less frequent updating with changes to inflation and fine distributions, which almost exclusively increase with time. We note though that partial information schemes might suffer from the introduction of undesirable ambiguity about fairness of punishment administered across violators. Future work is required to empirically evaluate the importance of these concerns against the benefits of increased deterrence.

meaningfully reduced failures to appear in court by clarifying information in summons forms. We thus believe that a logical next step for our research is to apply our "optimal punishment scheme" in a similar field context.

The remainder of the paper is structured as follows. In Section 2, we survey the related literature. Section 3 describes the experimental protocol of the main experiment and theoretical predictions. Section 4 presents the results of the main experiment. Section 5 presents the follow-up experiment, which replicates the results of the main experiment and expands on the mechanism driving subjects' behavior in the treatments with partial information. Section 6 concludes with a discussion of practical implications.

2 Related Literature

Our work relates to two strands of literature. The first one is concerned with measuring the prevalence and determinants of lying behavior in laboratory experiments (Fischbacher and Follmi-Heusi (2013), Gneezy, Rockenbach, and Serra-Garcia (2013) and references mentioned there).[6] Different from this literature, our focus is on mechanisms that prevent lying rather than on measuring the extent of lying per se.

The second strand, motivated by theoretical analyses of crime and law enforcement (see the classic model from Becker (1968)), is an active and fascinating experimental literature which investigates interventions and their effectiveness at reducing undesirable behavior in the lab. Engel (2016) provides a comprehensive survey of this research.[7] In particular, experiments have documented that more severe punishments are more successful at deterring criminal activity (Engel and Nagin (2015) and references mentioned therein), compared the deterrence effects of increasing monitoring probability versus the severity of punishment (Nagin and Pogarsky (2003), Friesen (2012), and Feess et al. (2015)), and explored how effective social norms are at deterring undesirable behavior (Dwenger et al. (2016), Casagrande et al. (2015)). As far as we know, our paper is the first to compare the effectiveness of different information structures describing punishments while holding fixed the monitoring probability and the severity (expected value) of the punishment.

The two most closely related papers to ours are DeAngelo and Charness (2012) and Salmon and Shniderman (2019). DeAngelo and Charness (2012) consider how varying jointly monitoring probabilities and fines affect deterrence rates, and, specifically, focus on

[6] See also Tergiman and Villeval (2019) for an experimental study of effects of reputation on lying behavior in the markets. In addition, Erat and Gneezy (2011) investigate a different type of lie, called 'white lie', which may benefit the person on the receiving end of the lie.

[7] See also Horne and Rauhut (2011) who evaluate the strength and weaknesses of the experimental approach in studying crime and law enforcement questions.

the link between preferences for punishment regime and compliance rates. The authors find that violations are less likely when the expected cost of violation is higher and when there is uncertainty about which regime is implemented. Contrary to our paper, however, the authors do not study partial information regimes and focus on the settings in which probabilities of each regime are common knowledge among participants.

Salmon and Shniderman (2019) conduct a tax compliance experiment to illustrate how individuals respond to ambiguous punishment probabilities and, in particular, how they respond to shifts in ambiguous versus known probabilities. They find that when probabilities are known and shift, the standard model works well to explain the behavioral response. Whereas when the probabilities are ambiguous and shift, the behavioral response is minimal. Related experiments have sought to infer the probabilities of being caught as agents' perceive them (Bebchuk and Kaplow (1992)).[8] However, no previous work has systematically investigated how the information revealed about the fine distribution of a punishment scheme influences deterrence behavior, an important gap in the literature, which we attempt to fill.

3 Main Experiment

3.1 Experimental Protocol

All experimental sessions were conducted at the Experimental Economics Laboratory at the University of California in San Diego between March 2019 and June 2019.[9] Since our experiment was short (it took approximately 5 minutes to complete), in lieu of recruiting subjects exclusively for our experiment, we asked other experimenters to add it at the end of the experimental session as an additional task.[10] Our instructions were very clear about the fact that the task performed by subjects in this last part of the experiment has nothing to do with the previous parts, and that their payment for the two tasks were independently determined. Overall, 424 students from the general population of UCSD participated in our experimental sessions. The experiment was programmed in O-Tree (Chen, Schonger, Wichens (2016)).

[8]See also theoretical model of Calford and DeAngelo (2020), in which agencies who wish to minimize criminal activity should reveal their resource allocation if criminals are uncertainty seeking and shroud their allocation if criminals are uncertainty averse. The authors supplement theoretical analysis with experimental evidence largely consistent with the theoretical predictions.

[9]We thank the researchers at UCSD Experimental Economics lab for their generosity in allowing us to run these sessions.

[10]We have made sure that subjects participated in no more than one experimental session.

Motivated by a variety of punishment schemes used in law enforcement, we conducted five different treatments. In all treatments, a subject is allocated one of five cards labeled with numbers 1 through 5, selected at random. The task is to report the number on the card one receives. If a subject reports a number x, then she earns $\$2x$. The treatments differ by the presence of monitoring and the fine that a subject incurs if she is caught lying. In the BASELINE treatment, there is no monitoring and no fines, i.e., subjects simply report their card number and collect their payments. In the remaining four treatments, there is a 20% chance that a subject is audited and punished if she lied, i.e., reported a number different from the number specified on her allocated card.

In the FIXED treatment, a subject who is caught misreporting her card number pays a fine of \$5. In the RANDOM treatment, the fine is either \$3 or \$7 with equal chance. In the MINIMUM treatment, the fine is at least \$3, and, finally, in the MAXIMUM treatment, the fine is at most \$7. In actuality, for the MINIMUM and MAXIMUM treatments, we use the same distribution of fines as in the RANDOM treatment, i.e., the fine is either \$3 or \$7 with equal chance, but subjects do not know this fact. In all cases, the fines are subtracted from the earnings that are based on the reported number.[11]

The experiment was conducted using the strategy method, i.e., subjects had to submit a number for each of the five possible cards. Then, to determine their payment, the computer randomly selected one of the five cards and calculated the subject's earnings based on the report provided for the selected card and the monitoring/punishment scheme specified by the treatment. The monitoring was implemented by a random draw performed by the computer for each subject individually.[12] The advantage of using the strategy method is that we are able to collect individual level data corresponding to subjects' choices for all five cards they may receive.[13]

Table 1 summarizes our experimental treatments and the number of participants. The instructions for all treatments are presented in Appendix 1.

[11]For instance, a subject in the FIXED treatment who received a card with number 3, reported number 4 and was caught lying earns $\$3 = (2 \cdot \$4) - \$5$.

[12]One might worry that the mere fact that subjects knew that the experimenter is 'observing' their choices, i.e., collects their reports for all cards, may impact one's desire to misrepresent the received card number. This concern motivated the majority of previous experiments studying lying behavior in the lab to adopt a design in which subjects privately roll the dice in a cubicle without anyone observing them and then report the number they rolled. However, this design does not naturally lend itself to an investigation of the effectiveness of various punishment schemes, since one needs to know the side of the rolled dice to be able to determine whether a subject lied or not. This was the primary reason we modified the design of standard lying experiments. As we will show in the next section, this change did not affect the main empirical regularities we observed, which is re-assuring as behavior seems to be stable to variations in the experimental protocol.

[13]See Gneezy, Rockenbach, and Serra-Garcia (2013) who also use the strategy method to elicit individual level tendency to lie in a laboratory experiment.

Table 1: Experimental Treatments

Treatment	Monitoring probability	Fine if caught lying	# of subjects
BASELINE	no monitoring	no fine	84 subjects
FIXED	20% known	\$5 for sure	80 subjects
RANDOM	20% known	\$3 or \$7 with equal chance	88 subjects
MINIMUM	20% known	at least \$3	92 subjects
MAXIMUM	20% known	at most \$7	80 subjects

In addition, we have access to individual characteristics of participants collected in the experiment conducted before ours. These controls include an IQ measure, risk attitudes, and overconfidence. To measure IQ, subjects were asked to solve six Raven matrices and received 50 cents for each correctly solved puzzle. Subjects' overconfidence was measured using two related characteristics, over-estimation and over-placement (we used the procedure similar to Chapman et al. (2019)).[14] For measuring over-estimation, subjects were asked to estimate how many of the six Raven puzzles they solved correctly; the correct answer was rewarded by 50 cents. For measuring over-placement, subjects were asked to rank themselves in terms of how many correct puzzles they solved relative to 75 other UCSD students who completed this task before; the correct answer was rewarded by 50 cents. The risk attitudes were measured using two investment tasks, in each of which subjects were endowed with 200 points (worth a total of \$2), any portion of which they could choose to invest in a risky project. In the first investment task, the risky project was successful 50% of the time and had a return of 2.5 points for each point invested in it, while in the second investment task the risky project was successful 40% of the times and returned 3 points for each point invested in it. Points not invested in the risky project had a return of 1 to 1 point. One of the two investment tasks was randomly chosen for payment. We administered this task twice with two sets of parameters as described above in order to use the econometric technique ORIV developed by Gillen, Snowberg, and Yariv (2018) to reduce measurement error.[15]

[14] Over-estimation compares a subject's actual performance with her estimate of it. Over-placement talks about subjects' perceived performance relative to other participants in the specified group.

[15] The investment task was developed by Gneezy and Potters (1997) and is one of the methods used in experimental literature to elicit subjects' attitudes towards risk (see the survey of experimental methods for various risk elicitation methods by Charness, Gneezy, and Imas (2013)).

3.2 Theoretical Predictions

To guide our experimental investigation, it is helpful to consider predictions made by various behavioral theories. We discuss separately the BASELINE treatment with no monitoring, treatments with full information about fines (FIXED and RANDOM), and treatments with partial information about fines (MINIMUM and MAXIMUM). We focus on how effective each punishment scheme is at preventing lying behavior and summarize this discussion as a series of hypotheses.

Treatment with no monitoring. The behavior in the BASELINE treatment is straightforward. Any model in which subjects' preferences are determined solely by the payoffs they earn in the experiment rather than psychological considerations such as self-image will predict that subjects should report number five for all cards in the BASELINE treatment.

Hypothesis 1: *A subject whose preferences are determined by payoffs earned in the experiment is expected to report number five for every card in the* BASELINE *treatment.*

Treatments with full information about fines. This group includes the FIXED and the RANDOM treatments. A subject with preferences monotonic in payoffs is supposed to either report their true card number or to report number five, i.e., lie to the fullest extent. This follows from the fact that the probability of monitoring and the size of the fine do not depend on the extent of lying but only on the mere fact of lying.[16] In particular, models with monotonic preferences will not be able to accommodate self-sabotage behavior of two types: reporting numbers below the card number (such that their net total will be less than if they reported their card number), or reporting numbers above the card number but below five (such that their net total, if fined, would be less than if they reported five).

The comparison between the BASELINE and the FIXED or the RANDOM treatments depends on the model that guides one's behavior. We first consider the *Expected Utility* theory. A subject's risk attitude determines whether she would benefit from lying in the FIXED or the RANDOM treatments. Denote by $u(\cdot)$ subject's utility from monetary payments. Then, for the FIXED treatment, if there exists a value of $x \in \{1, 2, 3, 4\}$ such that

$$u(\$x) > 0.8 \cdot u(\$10) + 0.2 \cdot u(\$5)$$

then this subject will report card values truthfully for all cards that satisfy the above inequality and will report number five for all other cards. For the RANDOM treatment, the

[16] Our experimental instructions were very clear about the fact that the probability of being monitored and the size of the fine is independent of both received and reported card numbers (see Appendix 1).

relevant inequality is

$$u(\$x) > 0.8 \cdot u(\$10) + 0.2 \cdot \big[0.5 \cdot u(\$3) + 0.5 \cdot u(\$7)\big]$$

Notice that a risk-neutral subject would lie for all cards in both the FIXED and the RANDOM treatments. That is, we expect to observe no difference in behavior between the BASELINE, the FIXED, and the RANDOM treatments provided that subjects are profit-maximizing. However, a risk-averse subject is expected to lie at least as much in the FIXED compared with the RANDOM treatment since the latter represents a mean-preserving spread of the former. Further, a subject with a sufficiently concave utility function is expected to lie strictly less in the FIXED compared with BASELINE treatment and strictly less in the RANDOM compared to FIXED treatment.

We now consider the *Prospect Theory* by Kahneman and Tversky (1979). Since subjects cannot make losses in our experiment, Prospect Theory can make potentially different predictions from those of Expected Utility theory only if a subject evaluates gains and losses relative to a non-zero reference point. One natural reference point could be the expected payoff of lying, which is the same for all cards and is equal to $0.2 \cdot \$5 + 0.8 \cdot \$10 = \$9$. Denote by $u(\cdot)$ and $\lambda \cdot u(\cdot)$ the two parts of the utility function that a subject uses to evaluate risky alternatives, where $u(\cdot)$ is applied for payoffs above and $\lambda u(\cdot)$ for payoffs below the reference point and $\lambda > 1$. Then a subject is predicted to tell the truth for card x in the FIXED treatment if

$$\lambda u(\$2x) > 0.2 \cdot \lambda u(\$5) + 0.8 \cdot u(\$10)$$

and report five otherwise. Similarly, a subject is expected to report card x truthfully in the RANDOM treatment if

$$\lambda u(\$2x) > 0.2 \cdot \lambda \left[0.5 \cdot u(\$3) + 0.5 \cdot u(\$7)\right] + 0.8 \cdot u(\$10)$$

and lie otherwise. Similar to Expected Utility theory predictions, a subject who is loss averse for payoffs below the reference point and risk-averse above it will lie weakly less in the RANDOM compared with the FIXED treatment, because of concavity of function $u(\cdot)$. The next hypothesis summarizes this discussion.

Hypothesis 2: *A subject who obeys the postulates of Expected Utility theory and has linear or concave utility function will exhibit the following ranking of lying propensities across treatments for each card value:* BASELINE $\geq$ FIXED $\geq$ RANDOM*, where lying means reporting number five. The same prediction holds for a subject who acts according to Prospect Theory with a reference point determined by the expected payoff of lying. The curvature of*

the utility function determines whether any of these inequalities is strict.

Treatments with partial information about fines. This group includes the MINIMUM and the MAXIMUM treatments. In what follows we will impose two mild restrictions on subjects' beliefs in these treatments: (a) fines are non-negative and (b) subjects cannot make losses in the experiment.[17] We call beliefs reasonable if they satisfy both (a) and (b).

Behavior in the treatments with partial information about fines depends on subjects' beliefs, which are purposely not controlled or induced in our experiment. This is what makes formulating hypotheses about the MINIMUM and the MAXIMUM treatments tricky. In particular, subjects might entertain several beliefs about possible fines they would face if caught lying and, thus, condition their behavior on the subjective distribution over these sets of beliefs. Note that the sets of all reasonable beliefs, defined by conditions (a) and (b) above, are quite different in the two treatments: in the MAXIMUM treatment reasonable beliefs lie between \$0 and \$7, while they lie between \$3 and \$10 in the MINIMUM treatment. Thus, the average fine over all reasonable beliefs in the MINIMUM treatment (\$6.5) is higher than the one in the MAXIMUM treatment (\$3.5), which could be one reason why the MINIMUM treatment appears more effective at deterring lying compared with the MAXIMUM treatment. The same prediction would follow if subjects believe that they would face the worst conceivable fine among those that are implied by reasonable beliefs in the two treatments, i.e., a fine of \$10 in the MINIMUM treatment and a fine of \$7 in the MAXIMUM treatment. However, one can construct subjective probabilities over subsets of reasonable beliefs which would result in the opposite prediction, i.e., MINIMUM treatment being less effective at deterring lying than the MAXIMUM treatment.[18] Ultimately, it is an empirical question of which of the two partial information treatments deter lying more successfully, as well as how behavior in these two treatments fares against that observed under treatments

[17] Restriction (a) is based on the common interpretation of the word fine, which we believe is reasonable. Restriction (b) hinges on the idea that it is common knowledge among participants in the laboratory experiments that they cannot make losses by participating in the experiment.

[18] For instance, a subject who believes that she would face fines between \$3 and \$5 in the MINIMUM treatment and fines between \$5 and \$7 in the MAXIMUM treatment might appear more truthful in the MAXIMUM compared to MINIMUM treatment.

with full information about fines.[19,20,21]

Hypothesis 3: *A subject who believes she will face the highest reasonable fine if caught lying or the average reasonable fine in the two partial information treatments is predicted to lie more often in the* MAXIMUM *than in the* MINIMUM *treatment.*

However, the ranking of the two partial information treatments would be reversed if behavior is shaped by the anchoring heuristic. Specifically, we consider the possibility of subjects anchoring on the value specified in the punishment scheme. Following Tversky and Kahneman (1974), subjects might start from the most "accessible" number - equal to the initial value they see - \$3 in the MINIMUM treatment and \$7 in the MAXIMUM treatment and then adjust their belief from here. Oftentimes, the adjustment is in the right direction but not enough to remove the initial bias towards the anchor. Anchoring would therefore predict that subjects perceive fines to be lower in the MINIMUM treatment compared with the MAXIMUM treatment, and it follows that subjects would be expected to lie more in the former.

Hypothesis 4: *A subject who anchors to the fine number specified by the punishment scheme is expected to lie more often in the* MINIMUM *than in the* MAXIMUM *treatment.*

4 Results of the Main Experiment

We report our results in the following order. First, we validate the use of the strategy method by comparing behavior observed in our BASELINE treatment with that documented

[19]Note that without further information about subjective beliefs in the treatments with partial information, the standard models of ambiguity do not make a definitive prediction about effectiveness of the MINIMUM versus MAXIMUM treatment (see Gilboa and Schmeidler (1989) and survey of Machina and Siniscalchi (2013)). Further, while there is a growing literature in decision theory on unawareness, which considers the case of unknown states and hence obviously unknown probabilities (see Heifetz, Meier, and Schipper (2006) and Karni and Viero (2013)), we feel that our setting is better characterized by ambiguity environment given that the set of reasonable beliefs is well defined but the probabilities of the states are unknown.

[20]The comparison between lying propensities in the two partial information treatments and the other treatments depends on the utility curvature of a subject if we consider the Expected Utility Theory. For instance, a profit-maximizing subject with reasonable beliefs is expected to lie to the fullest extent in all five treatments, with potentially one exception for the card value four in the MINIMUM treatment, in which telling the truth about this card and lying results in the same expected payoff. On the other hand, a sufficiently risk-averse subject who believes she will face the worst reasonable fine in both partial information treatments if caught lying is expected to have the following ranking of lying propensities across treatments for all cards: BASELINE $\geq$ FIXED $\geq$ RANDOM $\geq$ MAXIMUM $\geq$ MINIMUM.

[21]Unlike the first two hypotheses, Hypothesis 3 does not illuminate the mechanism which drives behavior observed in the treatments with partial information. We will dive into exploring this mechanism in the follow-up experiment, which we describe and analyze in Section 5.

in previous literature. Second, we investigate the aggregate effects of monitoring and punishments and run the horse-race between treatments to determine the most effective one. Third, we examine responses to the various punishment schemes using individual-level data and classify subjects into types based on their choice profiles.

4.1 Approach to Data Analysis

All the statistical tests are performed using regression analysis. Specifically, we regress the outcome of interest, e.g. the indicator of lying or reporting number five, on a constant and an indicator for the considered treatment. We cluster observations at the individual level when we compare average treatment efficacy at deterring unwanted behavior across *all* cards; this is done to account for interdependencies of observations that come from the same subjects within a session. No clustering is used when we consider reports generated for a specific card since each subject has just one report per card. We report the p-value associated with the null hypothesis that two groups have the same average behavior.

To explore determinants of individual behavior across treatments, we conduct additional regression analysis, in which we control for individual characteristics of subjects, such as risk, overconfidence and measure of their IQ. Specifically, we run linear probability models and utilize the obviously related instrumental variables (ORIV) method developed by Gillen et al. (2018). This method deals with measurement error problems, which are inherent in any elicitation procedure including eliciting subjects' risk attitudes. ORIV's identifying assumptions conform to the standard conditions for identification when applying instrumental variables to the classical measurement error problem. First, the linear regression model must satisfy the usual identification conditions in the absence of measurement error. Second, the measurement error in each elicitation must have mean zero conditional on the other elicitation.[22]

4.2 Baseline treatment and the previous literature

In order to study the effectiveness of monitoring and punishment schemes, we had to modify the standard protocol used in the lying experiments literature. In this section, we compare behavior observed in our BASELINE treatment with key findings reported in Fischbacher and Follmi-Heusi (2013), FFH hereafter, a paper which inspired hundreds of studies on lying behavior. The design adopted by FFH and nearly all papers in this literature following FFH, is best known as the dice-in-a-cup design. In this set-up, participants shake a six-sided

[22]The instruments' relevance conditions are trivially satisfied since each elicitation provides a noisy measurement of the same underlying feature.

die in a cup and report the number they receive. Subjects earn higher payoffs for reporting higher numbers except for number six, for which they earn zero.[23] Researchers then study lying in aggregate by looking at how the realized distribution of reported numbers differs from the distribution one would expect assuming a fair die.

The underlying premise of the dice-in-a-cup design is that image issues might lessen otherwise present lying behavior if experimenters were to observe subjects' actual roll of a die. This premise, while natural, has never been explicitly tested, which is a surprise given that the aggregate data in the dice-in-a-cup experiments does not allow examination of individual behavior crucial for the investigation of lying phenomena. In contrast, our experiment utilizes the strategy method in order to allow for collection of rich individual-level data. The strategy method has been used extensively in laboratory experiments in the past few decades and has delivered important substantive results in many tasks. Brandts and Charness (2011) compare the use of the strategy method with the direct response method and find that by and large there is no convincing evidence which suggests that the strategy method delivers systematically different results across a variety of tasks and individual decision problems.[24] While this cross-game evidence is re-assuring, one needs to establish that the strategy method in this particular paradigm does not alter lying behavior when compared with the direct response method.

FFH document three key empirical regularities. First, some people seem to be honest and truthful in reporting the number they roll.[25] Second, a greater-than-expected fraction of the payoff maximizing number is reported (number five), indicating that many people lie. Third, a greater-than-expected fraction of people report four, which gives them the second-highest possible payoff.[26]

Data in our BASELINE treatment exhibit the same three regularities. First, we observe that 23% of subjects report the actual card number for all five cards. Second, the majority of subjects (64%) report the number five for all five cards, i.e., the number that delivers the highest payoff. Third, 4% of subjects report payoff-maximizing numbers (number five) sometimes but not always.[27] We reject the null that proportions of any of the described

[23]Thus, the payoff-maximizing behavior is to report number five.

[24]Specifically, Brandts and Charness (2011) compare twenty-nine existing comparisons among which sixteen find no difference, four do find differences, and nine comparisons find mixed evidence. Importantly, in no case do the authors find that a treatment effect found with the strategy method is not observed with the direct-response method.

[25]One would never be able to detect with certainty whether or not people are reporting truthfully, since no one observes their roll of a die. Thus, this conclusion is based on the observation that a non-negligent fraction of the lowest numbers are reported.

[26]This last observation is consistent with the idea of self-image, according to which people lie but not to the fullest in order to 'preserve some self-dignity'.

[27]In particular, subjects report number four, which gives them the second-highest payoff, non-negligible

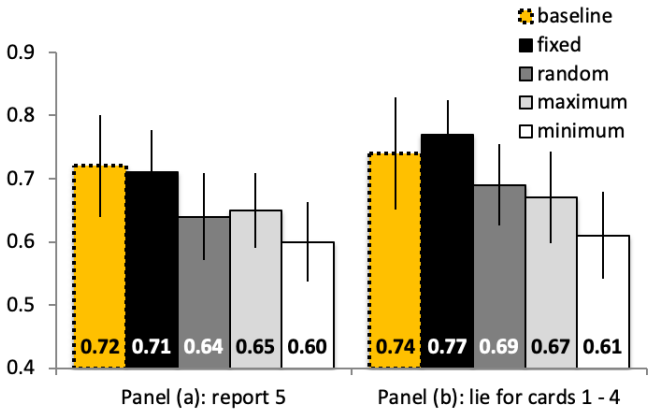
above types in our data is zero ($p < 0.001$ for the first two groups and $p = 0.0832$ for the last group of subjects). We conclude that lying behavior in the lab is not very sensitive to modifications of the experimental protocol. In other words, there are no losses associated with moving from the dice-in-the cup design to the strategy method, but there are significant gains since it allows us to collect richer individual-level data.

Observation 1: *Observing individual-level choices of subjects using the strategy method in the lying experiment produces the same aggregate lying behavior as the commonly used dice-in-the-cup paradigm, with the additional benefit of capturing individual-level data.*

4.3 Aggregate Effects of Monitoring and Punishments

We start by looking at the aggregate frequencies of lying across treatments. Figure 2 presents two different statistics: panel (a) depicts the frequency of reporting five averaged across all cards and panel (b) reports lying propensities for all cards except for Card5, where a lie is defined as reporting a number different from the card number.

Figure 2: Aggregate Frequencies of Lying



Notes: Panel (a) reports frequency of reporting number five averaged across all cards, by treatment. Panel (b) reports average lying propensities across Card1, Card2, Card3, and Card4, where a lie is defined as reporting a number different from the true card value. Error bars are the 95% confidence intervals based on robust standard errors obtained by clustering at the subject level to account of interdependencies of observations that come from same subjects.

Figure 2 shows that the MINIMUM scheme emerges as the most effective one at deterring unwanted behavior among all schemes considered in our experiment. Statistical analysis

fraction of time: 3.57% when they are allocated Card1, 7.14% when they get Card2 and 3.57% when they get Card3.

confirms that both the frequency of reporting number five and the frequency of lying for cards other than Card5 are significantly lower in the MINIMUM treatment as compared with the BASELINE treatment with $p = 0.021$ and $p = 0.030$, respectively. At the same time, the other punishment schemes appear to be less effective at the aggregate level ($p > 0.10$ for pairwise comparisons between the BASELINE and the other punishment schemes). This result suggests that investing resources in monitoring crime does not always pay off and it depends on the way punishment is presented. For instance, lying frequencies are statistically not different between the FIXED and the BASELINE treatments, suggesting that the policy maker would be better off by not implementing any monitoring or punishment compared with wasting resources to implement the fixed fine scheme.

Comparing the effectiveness of different punishment schemes conditional on monitoring, we find that both measures of aggregate behavior, i.e., reporting number five and lying for cards 1 thru 4, are the lowest in the MINIMUM treatment. Statistical analysis reveals that while the MINIMUM scheme is significantly better at preventing lying relative to the FIXED scheme ($p = 0.021$ and $p = 0.001$ for each of the two measures respectively), the aggregate levels of deterrence reached in the MAXIMUM and the RANDOM schemes are comparable with those in the MINIMUM treatment.[28]

Aggregate data masks important differences in behavior across treatments for different opportunity costs of lying as captured by the different card values. Figure 3 makes this point and depicts lying propensities for each card separately.

Figure 3 suggests several insights. First, while lying frequencies are quite stable across cards in the BASELINE treatment, they tend to (weakly) decline as the opportunity cost of lying decreases from Card1 to Card4 in all other treatments with monitoring and punishments.[29] Second, none of the punishments are effective at deterring lying for Card1, for which the incentives to lie are the highest.[30] On the contrary, when incentives to lie are relatively small, as is the case for Card4, all punishment schemes are equally effective at reducing lying as compared with the BASELINE treatment without monitoring.[31]

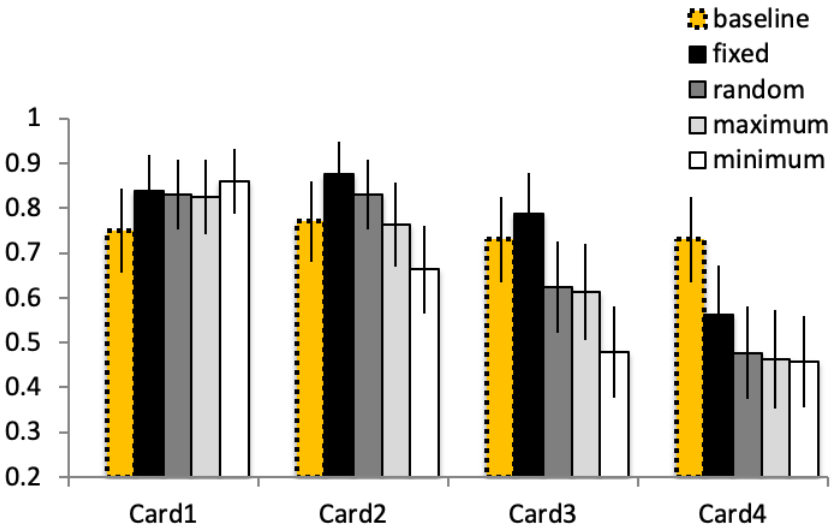
[28]For MINIMUM vs. RANDOM the corresponding numbers are $p = 0.392$ and $p = 0.104$. For MINIMUM vs. MAXIMUM the corresponding numbers are $p = 0.311$ and $p = 0.397$.

[29]For the BASELINE treatment, lying propensities are not significantly different at the standard 5% level for any two cards depicted in Figure 3 except for Card2 vs Card3 and Card2 vs Card4 ($p = 0.045$ in both cases). For the FIXED treatment, we obtain $p = 0.183$ for Card1 vs Card2, $p = 0.129$ for Card2 vs Card3, and $p = 0.000$ for Card3 vs Card4. For the RANDOM treatment, the p-values are $p = 1.000$, $p = 0.000$, and $p = 0.004$. For the MAXIMUM treatment, the p-values are $p = 0.025$, $p = 0.010$, and $p = 0.002$, and, finally, for the MINIMUM treatments the p-values are $p = 0.000$, $p = 0.000$, and $p = 0.621$. In all the regressions we cluster observations by subject to account for interdependencies of reports that come from the same individual.

[30]For any pair of treatments, lying frequencies for Card1 are not significantly different at the standard 5% level.

[31]For Card4, we obtain $p = 0.029$ for the BASELINE vs FIXED, $p = 0.001$ for the BASELINE vs the RANDOM,

Figure 3: Lying Frequencies for Each Card



Notes: Frequencies of lying are plotted for each card separately with the error bars depicting the 95% confidence intervals, calculated based on standard errors of the means.

Finally, the most separation between treatments comes from reports for intermediate cards, i.e., Card2 and Card3. In these two cases, the MINIMUM treatment is the most effective at deterring lying both as compared with the BASELINE treatment and as compared with other punishment schemes. Regression analysis presented in the first two columns of Table 2 confirms what we see in Figure 3. The MINIMUM treatment reduces lying for both Card2 and Card3 and this reduction is both large in magnitude and significant.[32] The behavior in response to these intermediate cards hints that the effectiveness of the MINIMUM treatment is related to beliefs about whether the transgression is profitable or not. For example, if one lies for Card3 and reports five instead, they stand to gain \$4 but lose *at least* \$3, making it fairly unattractive to lie. In other words, the MINIMUM scheme operates by removing the potentially optimistic expectation that a transgression could be profitable if caught.

Relating these results to the hypotheses outlined in Section 3.2, we first note that consistent with Hypothesis 1, subjects report number five in more than 75% of cases irrespectively of the card they actually receive in the BASELINE treatment. There is, however, 23% of subjects who always report card numbers truthfully in the BASELINE treatment

$p < 0.001$ for the BASELINE vs MAXIMUM, and $p = 0.001$ for the BASELINE vs MINIMUM.

[32]The comparison between the MINIMUM and the FIXED or the RANDOM treatments is statistically significant at the 1% level, while it is significant at 10% level for MINIMUM versus MAXIMUM treatments.

despite the fact that there is no monitoring and monetary fines for lying. Second, our data is consistent with Hypothesis 2, which assert that subjects should lie weakly more often in the BASELINE compared with FIXED treatment and weakly more often in the FIXED compared with RANDOM treatment. Indeed, as Figure 3 and first two columns of Table 2 show, the propensity to lie is not significantly different at the standard 5% level between the FIXED and the RANDOM treatments for every single card. Third, our data shows clear support for Hypothesis 3, according to which among the two partial information treatments, the MINIMUM one is more effective at deterring lying than the MAXIMUM one, and refutes Hypothesis 4, which predicts the opposite relation.

Table 2: Main Experiment: Regression Analysis

	Dependent Variable: Indicator for				
	Reg. (1)	Reg. (2)	Reg. (3)	Reg. (4)	Reg. (5)
	Lie Card2	**Lie Card3**	**Type *52345***	**Type *55345***	**Type *55545***
Indicator RANDOM	-0.02 (0.06)	-0.12* (0.07)	-0.01 (0.02)	0.23** (0.11)	-0.08 (0.10)
Indicator MAXIMUM	-0.08 (0.06)	-0.14* (0.07)	0.04 (0.04)	0.22* (0.11)	-0.23** (0.10)
Indicator MINIMUM	-0.21*** (0.06)	-0.31*** (0.07)	0.34*** (0.07)	0.04 (0.10)	-0.31*** (0.09)
Constant	0.66*** (0.12)	0.55*** (0.14)	0.04 (0.11)	0.29 (0.21)	-0.16 (0.19)
Individual Controls					
Risk attitudes	0.00 (0.00)	0.00** (0.00)	0.00 (0.00)	0.00 (0.00)	0.00** (0.00)
IQ measure	0.02 (0.02)	0.01 (0.02)	0.00 (0.02)	0.02 (0.03)	0.08** (0.03)
Overprecision	0.02 (0.02)	-0.01 (0.03)	-0.01 (0.03)	0.04 (0.04)	0.02 (0.04)
Overestimation	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
# of observations	$n = 340$	$n = 340$	$n = 163$	$n = 163$	$n = 163$
adjusted R-squared	0.0464	0.0598	0.2276	0.0669	0.1212
Sample	All	All	Occasional Over-reporters	Occasional Over-reporters	Occasional Over-reporters
Tests of Coefficients					
MINIMUM = FIXED	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.718$	$p < 0.001$
MINIMUM = RANDOM	$p = 0.003$	$p = 0.008$	$p < 0.001$	$p = 0.055$	$p = 0.009$
MAXIMUM = FIXED	$p = 0.162$	$p = 0.059$	$p = 0.268$	$p = 0.054$	$p = 0.020$
MAXIMUM = RANDOM	$p = 0.269$	$p = 0.811$	$p = 0.194$	$p = 0.882$	$p = 0.115$
MINIMUM = MAXIMUM	$p = 0.063$	$p = 0.023$	$p < 0.001$	$p = 0.088$	$p = 0.287$

Notes: We report the results of ORIV (linear probability model) estimations with fixed treatment being the base group (see for details Gillen et. al. (2018)). The individual controls include (a) risk attitudes measured by the fraction of the endowment invested in the risky project, where higher investment indicates smaller degree of risk-aversion, (b) an IQ measured by the number of correctly solved Raven matrices, (c) the over-precision measured by the difference between the number of Raven matrices a subject thinks he solved correctly and the actual number he solved, and (d) over-placement measured by the difference between the predicted rank of a subject in a group of 100 undergraduate UCSD students and his actual rank. ***, ** and * indicates significance at the 1%, 5% and 10% level, respectively.

Observation 2: *Communicating the "minimum fine" of a fine distribution is most effective at deterring unwanted behavior both in aggregate and for cards with intermediate values of opportunity costs of lying.*

4.4 Effectiveness of punishment schemes at the individual level

To investigate individual level responses to various punishment schemes, we classify subjects into four mutually exclusive types based on their individual choice profiles:

1. ***Under-reporters*** are those who report a number strictly smaller than the card number for at least one card.
2. ***Honest reporters*** are those who truthfully report the card number for all five cards.
3. ***Occasional over-reporters*** are those who report numbers which are higher or equal to the received card number with at least one reported number being different than five.
4. ***Persistent over-reporters*** are those who report number five for all five cards.

Figure 4: Distribution of individual types

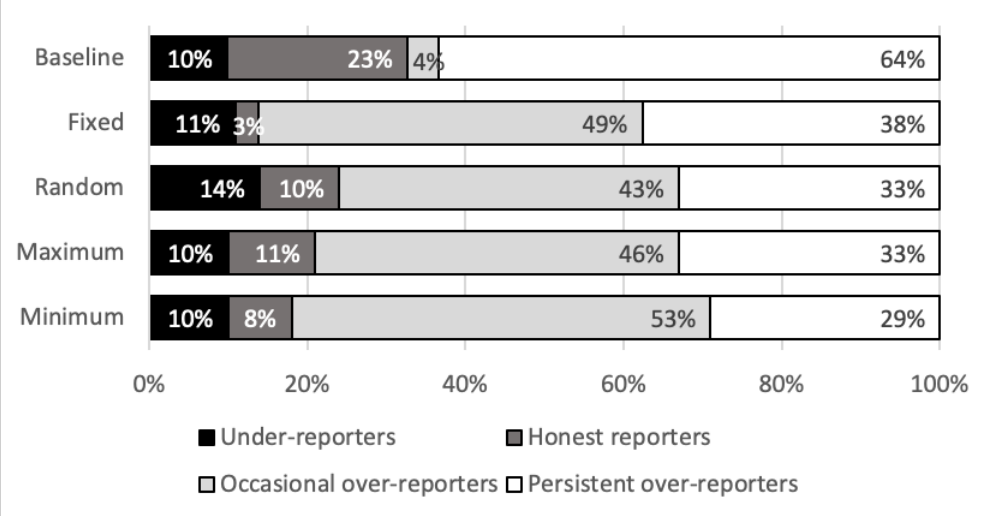
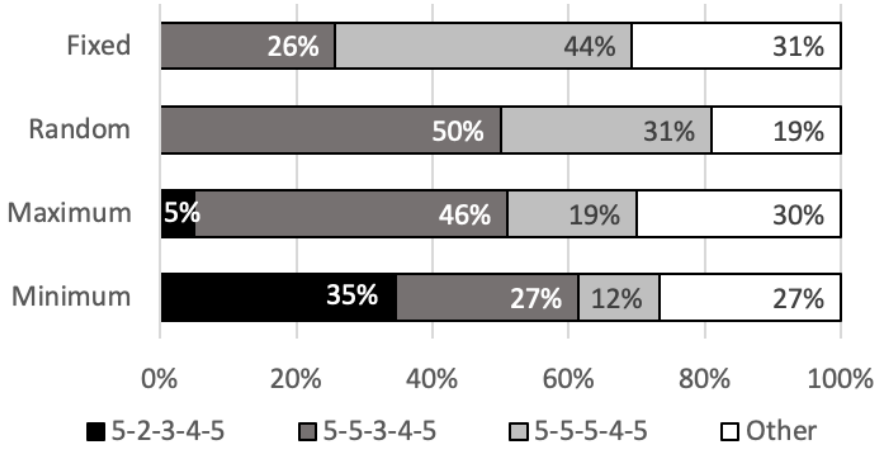


Figure 4 depicts the distribution of types in each treatment. Segmenting subjects into types is valuable because a policy maker who wishes to improve overall welfare is likely to be primarily concerned with reducing the proportion of persistent over-reporters in the population. Compared with the BASELINE treatment, all treatments with monitoring and punishment significantly reduce the subjects who always report number five. This reduction

is large in magnitude and statistically significant: it cuts this fraction from about two-thirds to a third of subjects or less depending on the punishment scheme ($p < 0.01$ in all pairwise comparisons). The reduction in the fraction of persistent over-reporters is accompanied by a large increase in the fraction of occasional over-reporters across all treatments, ranging from 4% in the BASELINE treatment to more than 40% in all other treatments.[33]

Interestingly, the introduction of monitoring and punishments significantly reduces the fraction of honest reporters subjects in the population from 23% in the BASELINE treatment to as little as 3% in the FIXED treatment[34] and about 10% in all other treatments.[35] This is consistent with the literature on how imposed incentives can backfire by crowding out intrinsic motivation (Gneezy, Meier and Rey-Biel (2011)). Finally, we note that the proportion of under-reporters remains stable across treatments and ranges between 10% and 15%.[36]

Figure 5: Types of Behavior among Occasional Over-reporters



Notes: Each category represents reports for each of the cards listed in the consecutive order. For instance, 5-2-3-4-5 category means that a subject reported 5 for Card1 and the true card value for all the remaining cards. Category other comprises all subjects who reported a number different from five but higher than the true card value at least once.

The distribution of types across different punishment schemes is quite stable as seen in

[33]We find $p < 0.01$ in all pairwise comparisons of fraction of occasional over-reporters between the BASELINE and all other treatments.

[34]Hence explaining our earlier result that a FIXED punishment scheme is not very effective overall.

[35]We obtain $p < 0.001$ for BASELINE vs FIXED, $p = 0.028$ for BASELINE vs RANDOM, $p = 0.005$ for BASELINE vs MINIMUM, and $p = 0.054$ for BASELINE vs MAXIMUM.

[36]We cannot reject the null that the fraction of under-reporters is the same in every pair of treatments with $p > 0.10$. We have no way to identify whether these are the subjects who did not understand the instructions or simply made a typing mistake in recording the numbers.

Figure 4 with no significant differences detected between any pair of treatments for any category. The only exception is the fraction of honest reporters, which is significantly smaller in the FIXED treatment compared with the RANDOM ($p = 0.043$) and with MAXIMUM ($p = 0.029$) treatments.

However, the similarity in population types across treatments hides differences between treatments within the category of occasional over-reporters. Note that this group encompasses two different individual choice profiles: subjects who lie for lower cards and report truthfully for higher cards, and subjects who lie but not to the fullest extent. Figure 5 presents the breakdown of occasional over-reporters into these categories distinguishing between subjects who lie only for Card1, lie for Card1 and Card2, lie for Card1, Card2, and Card3, and those who lie but not to the fullest extent possible (referred as other in the Figure 5). The last three columns in Table 2 report the results of the regression analysis conducted to detect differences in these sub-types across treatments.

Figure 5 and regressions reported in Table 2 reveal several regularities about occasional over-reporters. First, while the discussion in Section 3 suggests that subjects should only report number five or the true card number, we observe a non-negligible fraction of subjects who lie not to the fullest extent at least once. This is consistent with FFH observation of what they call 'incomplete liars'. Second, the MINIMUM treatment features the highest number of subjects who lie only for the lowest card, i.e., type 52345, among all punishment treatments (see Regression (3)). Moreover, the fraction of people who lie for all but one card (type 55545) is lower in the two partial information treatments MINIMUM and MAXIMUM compared with RANDOM and FIXED treatments as seen in Regression (5) and comparative tests of the estimated coefficients reported at the bottom of Table 2. In other words, while the fraction of subjects who lie occasionally is the same across punishment treatments, the MINIMUM treatment features the least amount of lying within this category.

Observation 3: *Monitoring and punishments of any kind cuts the fraction of persistent over-reporters by at least half. The punishment scheme that specifies the "minimum fine" is the most effective at reducing instances of lying within a category of occasional over-reporters.*

5 Mechanism Driving the Results

The results of our main experiment show that the treatment which uses the "minimum fine" outperforms other treatments in terms of deterring lying behavior, even when compared to another partial information treatment: the "maximum fine." In this section, we explore the

mechanism behind this result. As we discussed in Section 3.2, the relative effectiveness of the MINIMUM versus the MAXIMUM treatment depends on subjects' beliefs about the fine distributions, which are not induced nor controlled in the main experiment. The difference in beliefs across these two treatments could generate various predictions including the one in which the MINIMUM treatment outperforms the MAXIMUM one as well as the reverse. Ultimately, we need empirical evidence that links both beliefs and subjects' behavior in these two partial information treatments in order to identify the driving force behind results obtained in the main experiment.

To do that, we conducted a follow-up experiment focusing on the two partial information treatments. The new experiment serves two goals. First, we replicate the results of our main experiment to see how robust they are. Second, we elicit subjects' beliefs about fines in partial information treatments in an attempt to identify the main forces driving behavior.

5.1 Experimental Protocol of the Follow-up Experiment

The follow-up experiment consists of two treatments: MINBELIEFS and MAXBELIEFS with 96 and 94 subjects, respectively.[37] The two treatments are identical to the MINIMUM and the MAXIMUM treatments in the main experiment with the addition of two questions in which we elicited subjects' belief about the fine structure. The beliefs questions were presented to subjects in random order and administered at the end of the experiment before subjects' learned their payment for the cards task.

One of the questions elicited subjects' beliefs about the average fine one would pay if caught lying, and the second question asked subjects to state the fine they believe *they specifically* would pay if they were caught lying. Here is the exact formulation of the beliefs' questions in the MINBELIEFS (MAXBELIEFS) treatments:

Q1: *In Spring 2019, 90 (80) UCSD students participated in the experiment identical to the one you just finished. Just like in your experiment, with probability 20%, a subject's reported number was compared with the actual card number she received, and in case these were different a subject incurred a penalty of at least \$3 (at most \$7). In that experiment, what do you think was the average penalty of subjects who were selected (according to 20% rule) and found to misreport their card number? If your guess is*

[37]The new treatments were conducted in the same Experimental Economics Laboratory at UCSD in November 2019 at the end of the unrelated experiment. We made sure that no subjects had participated in the previous treatments.

within +/− 50 cents of the actual average penalty in that experiment, you will receive an additional $1."

Q2: *Think about the experiment you finished. What do you think would be your penalty if you were selected (according to 20% rule) and you reported a different number from the card number you received?"*

Few details of our beliefs' elicitation procedure deserve a discussion. First, we cannot incentivize the second question in which subjects report beliefs about their own potential fine. This suggests that we should take this measure with a grain of salt since it might be a noisy estimate of subjects' true beliefs. Second, while it would be interesting to elicit the whole distribution of beliefs that subjects' might consider, we opted for simpler and partial statistics about this distribution, i.e., the average fine and their own fine.[38]

Third, one interpretation of the difference between the answers subjects give to the two beliefs questions above is that this difference is related to subjects' attitudes towards ambiguity. Those who report their own fine to be strictly higher/lower than the average fine are ambiguity averse/seeking, while those who report the same answers are ambiguity neutral.[39] This interpretation is consistent with the model of smooth ambiguity preferences, which received much attention in the literature given its wide range of applications (Klibanoff, Marinacci, and Mukerji (2005), Seo (2009), Al-Najjar and De Castro (2014), Cerreia-Vioglio et. al. (2013), Klibanoff et. al. (2019), and Denti and Pomatto (2020)). According to this model, a decision maker evaluates acts using a two-fold expectation, with the first one computing expected utility for each individual probability measure separately and the second expectation aggregating over different probability measures through the lens of the ambiguity index function.[40] The shape of the ambiguity index function determines

[38]We chose to elicit the average rather than the median fine because the concept of average is commonly used, while median is not so much. This should not matter as long as subjects believe that the distribution of fines is symmetric.

[39]For the standard measure of ambiguity attitudes in the laboratory experiments see Halevy (2007).

[40]Formally, an act $f : \Omega \to X$, which maps states of the world to outcomes, is evaluated using the two-fold expectation

$$V(f) = \int_P \psi \left(\int_\Omega u(f) dp \right) d\mu(p)$$

where u stands for the utility function, ψ represents an ambiguity index, and μ stands for the belief over a set P of probabilities. In words, acts are first evaluated using their expected utility with respect to each probability measure p in the set P. Then, these expectations are averaged by means of a belief μ over probabilities and ψ which is an increasing transformation, i.e., the ambiguity index. This means that the decision-maker considers multiple probabilities provided that μ is not a singleton. When ambiguity index ψ is linear, the decision-maker displays ambiguity neutral attitudes. However, if ψ is not linear, then this model can accommodate both ambiguity averse and ambiguity seeking behaviors, which correspond to ψ being concave and convex, respectively. The ambiguity averse decision maker is someone who dislikes the uncertainty about beliefs over the beliefs.

a subject's attitude towards ambiguity, with concave functions corresponding to ambiguity aversion and capturing a subject's aversion to uncertainty about beliefs over fine distribution beliefs. This model of smooth ambiguity preferences provides a natural interpretation of our beliefs questions. When asked about one's own fine, a subject uses the ambiguity index function to evaluate various possibilities of the fines, which are unknown to the experimenter. However, when asked about the average fine, the ambiguity index function does not enter the calculation, and one simply reports the average over average fines computed based on each set of beliefs over possible fines in the corresponding treatment.

We note, however, that while our preferred interpretation of the spread of beliefs is the one described above, it could also encompass additional forces unrelated to ambiguity aversion. Among natural alternatives are subjects' beliefs regarding the differential treatment they may receive if caught lying relative to others or how fines are affected by differences between own and average lying frequencies. For these reasons, in the remainder of the analysis we refer to the difference between reported own and average fines as *beliefs spread*. We study the properties of the *beliefs spread* in the two treatments and explore how it relates to lying propensities in the cards task.

5.2 Results of the Follow-up Experiment

We start by comparing the aggregate results in the main and follow-up experiments. Figure 6 presents the same two aggregate statistics about lying propensity in the follow-up experiment as the one presented in Figure 2 for the main experiment. There is little difference in aggregate behavior between the main and the follow-up experiments conditional on punishment scheme.[41] Moreover, as in the main experiment, we find that at the aggregate level both the minBeliefs and the maxbeliefs are equally effective at deterring lying using either of the two measures, i.e., the frequency of reporting number five and the frequency of lying for all cards except for Card5.[42]

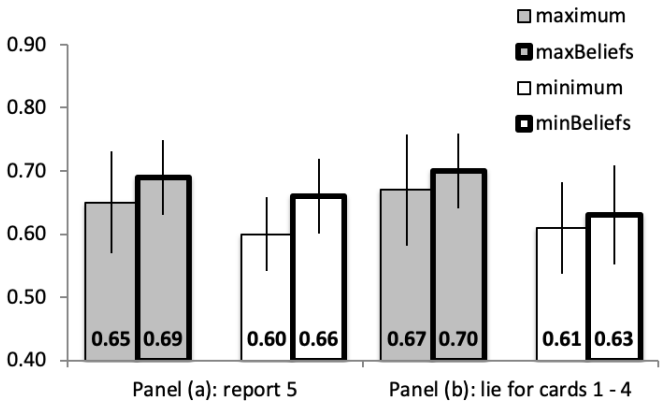
Next, we look at the distribution of types in the new treatments based on the card reports (see Table 3). The distribution of types across new treatments is stable and similar to the one observed in the main experiment with most subjects falling into either the occasional or persistent over-reporters category.

Table 4 reports the distribution of beliefs' types in the new treatments as well as sum-

[41] In Appendix 2, we present lying frequencies for each card separately and compare those to trends documented in the main experiment.

[42] We obtain $p = 0.523$ and $p = 0.164$ for the frequency of reporting five between and frequency of lying for cards 1 - 4, respectively.

Figure 6: Aggregate Frequencies of Lying in the Main and Follow-up Experiments



Notes: Panel (a) reports frequency of reporting number five averaged across all cards, by treatment. Panel (b) reports average lying propensities across Card1, Card2, Card3, and Card4, where a lie is defined as reporting a number different from the true card value. Error bars are the 95% confidence intervals based on robust standard errors obtained by clustering at the subject level.

Table 3: Distribution of Types in the Follow-up Experiment

	Under-reporters	Honest reporters	Occasional over-reporters	Persistent over-reporters
MINBELIEFS	5%	11%	50%	33%
MAXBELIEFS	7%	9%	46%	38%
	$p = 0.528$	$p = 0.501$	$p = 0.560$	$p = 0.478$

Notes: The p-values reported in the last row are obtained from the regression analysis.

mary statistics about the subjects' beliefs.[43] First of all, we note that average own fine in the two treatments (MINBELIEFS and MAXBELIEFS pooled together) is not significantly different from 5, which is the expected average fine used in all treatments of the main experiment.[44] Therefore, the effectiveness of partial information schemes compared with full information schemes does not come from subjects overestimating the expected fines they will have to pay if they are caught lying.

[43]Our program allowed subjects to enter any numbers they wish for both belief questions. As a result, 4 subjects have specified fines below \$3 in the MINBELIEFS treatment and 3 subjects have specified beliefs above \$7 in the MAXBELIEFS. In addition, there are 7 subjects in the MINBELIEFS treatment who specified that the average fine is above \$10, which is impossible by the design of the experiment. We have excluded these subjects from the analysis that follows, which leaves us with 91 subjects in the MAXBELIEFS treatment and 85 subjects in the MINBELIEFS treatment.

[44]The average own fine in both MINBELIEFS and MAXBELIEFS treatments pooled together is 4.93 with the standard error of 0.17. We cannot reject the null hypothesis that average own fine is equal to five with $p = 0.692$.

Second, the largest group in the population are subjects who believe that they would face a higher fine if caught lying relative to the average fine administered for the same violation; this group constitutes about half of subjects in both treatments ($p = 0.332$ and $p = 0.917$ for the MINBELIEFS and MAXBELIEFS treatments, respectively). The remaining subjects for the most part believe that they will face the same fine as the average person in the MINBELIEFS treatment and lower fine than the average in the MAXBELIEFS treatment. Consistent with the anchoring hypothesis, the average fine reported in the MAXBELIEFS treatment is significantly higher than that reported in the MINBELIEFS treatment ($p = 0.016$). This difference is mostly driven by subjects with same-as-average beliefs who believe that the fine would be on average \$1 more in the MAXBELIEFS than in the MINBELIEFS treatment. At the same time, subjects hold higher beliefs about their own fine in the MINBELIEFS compared to the MAXBELIEFS treatment ($p = 0.019$).[45]

Table 4: Distribution of Belief Types in the Follow-up Experiment

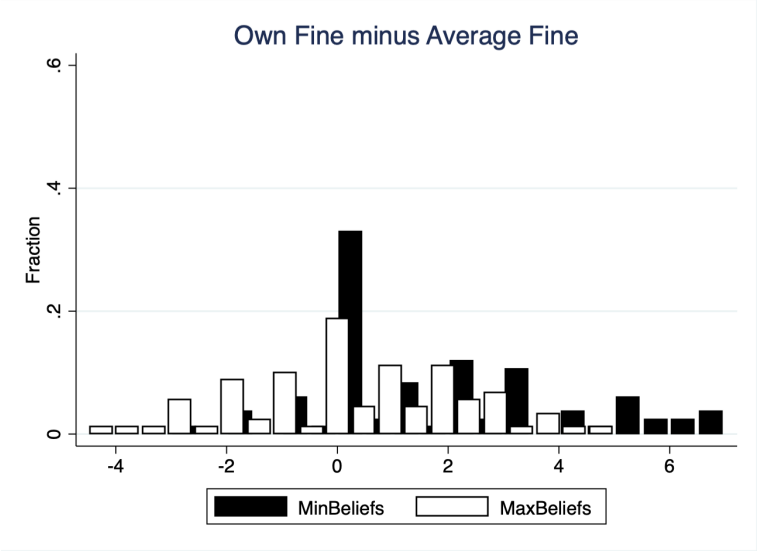
	Higher-than-average beliefs			Same-as-average beliefs		Lower-than-average beliefs			All Subjects beliefs	
	frac	ave	own	frac	ave/own	frac	ave	own	ave	own
MINBELIEFS	0.55	3.6	6.7	0.33	3.8	0.12	4.5	3.1	3.8	5.3
MAXBELIEFS	0.49	3.9	6.0	0.19	4.8	0.32	4.2	2.2	4.2	4.6
	$p = 0.441$	$p = 0.109$	$p = 0.035$	$p = 0.030$	$p = 0.007$	$p = 0.001$	$p = 0.460$	$p = 0.049$	$p = 0.016$	$p = 0.019$

Notes: We report the fraction of beliefs' types in the two treatments as well as mean average and own fines reported by each type. The last two columns list the mean average fine and mean own fine reported by all subjects in these two treatments. We exclude subjects who report unreasonable beliefs as defined in Footnote 43. The last row of the table reports the p-values comparing MINBELIEFS and MAXBELIEFS treatments from the regression analysis.

Figure 7 gives a fuller picture of subjects' beliefs by plotting the differences between own and average fines computed at the individual level. The picture shows that the distributions are quite different across the two treatments: MINBELIEFS treatment features a distribution which is skewed to the right relative to the MAXBELIEFS treatment, where the distribution is much closer to being symmetric around zero. Importantly, MINBELIEFS distribution features larger differences between own and average fines compared to MAXBELIEFS ($p < 0.001$). The same pattern holds if we condition on subjects who have higher-than-average beliefs only ($p = 0.001$). In other words, while the fraction of subjects with higher-than-average beliefs remains the same across two treatments, the MINBELIEFS treatment induces a greater *difference* between own and average fines as compared to the MAXBELIEFS treatment.

[45]Note that this result is despite the fact that we have removed a few outliers in the MINBELIEFS treatment, i.e., subjects who believe in fines above \$10.

Figure 7: Difference between Own Fines and Average Fines



Observation 4: *Beliefs regarding the average (own) fine are higher (lower) in the treatment in which a "maximum" fine is specified than in a treatment in which a "minimum" fine is specified. Furthermore, about a half of subjects report that they believe their own fine will be higher than that of an average person with the spread between these two beliefs being higher in the "minimum" than in the "maximum" treatment.*

We now turn to investigate the consistency between beliefs and actions in the card task. To this extent, we present in Table 5 the results of several regressions which look at different dimensions of lying in the cards task controlling for subjects' individual beliefs and individual characteristics including risk-attitude, IQ and overconfidence.

Specifically, in Panel A we present four different specifications of regressions, in all of which the dependent variable is the indicator for honest reporters, while in Panel B the dependent variable is an indicator for lying for Card2. In Regression (1) we show that among the two elicited beliefs, it is the belief about own rather than average fine which is correlated with being truthful in cards' reports. Regression (2) shows that there is a positive relationship between holding the belief that your own fine will be higher than average and being an honest type. Regression (3) shows that the spread of beliefs, i.e., own believed fine minus the average believed fine, also positively correlates with reporting all cards truthfully. Finally, Regression (4) observes the same relationship as Regression (3) within a subset of subjects who believe their own fine will be higher than average. In Appendix 2, we show that similar conclusions hold for regressions in which the dependent variables are persistent

over-reporters (Panel C) and the indicator for lying for Card3 (Panel D).

All the regressions paint a consistent picture. First of all, all measures of lying are correlated with own believed fines rather than average fines. Second, subjects who believe that their own fine will be higher than the average are less likely to lie[46]. Finally, the spread of beliefs is negatively correlated with lying behavior with a higher spread associated with less lying. This last result holds both when we look at all the subjects, and when we focus on the subset of only those who report higher than average beliefs for whom, by definition, this spread is positive.

Observation 5: *Beliefs about own fine rather than the fine faced by an average person in the population are correlated with behavior in the cards task. Moreover, lying in the cards task is negatively correlated with the belief that own fine will be higher than the average and the spread of beliefs, i.e., the difference between the own reported fine and the average one.*

6 Conclusion

In this paper, we asked whether organizations could reduce undesirable behavior by more strategically revealing information about anticipated fines. Specifically, we considered a situation in which an organization may have a constrained monitoring probability (e.g. the size of their compliance team only allows monitoring a fraction of the population at any given time) and the fine range used to punish violations is moderate (such that some violators will view the value of transgressing as being at least equal to or greater than the value of the fine). One of the few remaining tools at their disposal is therefore the information they reveal about the fine distribution. While we focus on the managerial domain, the trade-offs captured by our environment speak more generally to the variety of compliance and enforcement problems studied in public economics.

Given these constraints, the goal of the paper was to investigate the efficacy of various information structures at deterring unwanted behavior ("crime" in the lab, as captured by lying in our experimental set-up) and to uncover the mechanism underlying the behavioral results.

From a methodological point of view, we make two contributions. First, our experiment applies a well-known technique, the strategy method, to the classic lying paradigm, for which this method has not been used up to now. Our experimental results show that eliciting behavior at the individual level without anonymity produces the same aggregate

[46]Since the question about own expected fine was administered after the task and could not be incentivized, we cannot rule out the possibility that subjects were engaging in post hoc justification of their choices.

Table 5: Beliefs and Behavior in Follow-Up Experiment

Panel A	**Dependent Variable: Indicator for Honest Reporters**			
	Reg. (1)	Reg. (2)	Reg. (3)	Reg. (4)
Own Fine	0.03** (0.01)			
Ave Fine	-0.02 (0.02)			
Indicator Higher-than-Average Beliefs		0.14*** (0.04)		
Own Fine − Ave Fine			0.03** (0.01)	0.04* (0.02)
Indicator MINBELIEFS	-0.01 (0.06)	0.02 (0.04)	-0.02 (0.04)	-0.01 (0.08)
Constant	0.24 (0.15)	0.22** (0.11)	0.26** (0.11)	0.33 (0.22)
# of obs	131	176	176	92
adjusted R-sq	0.0747	0.0785	0.0781	0.1403
sample	higher-than-ave or lower-than-ave	all	all	higher-than-ave
Panel B	**Dependent Variable: Indicator for Lie Card 2**			
	Reg. (4)	Reg. (5)	Reg. (6)	Reg. (7)
Own Fine	-0.04** (0.02)			
Ave Fine	0.04 (0.03)			
Indicator Higher-than-Average Beliefs		-0.17*** (0.06)		
Own Fine − Ave Fine			-0.05*** (0.02)	-0.12*** (0.03)
Indicator MINBELIEFS	-0.10 (0.07)	-0.11* (0.06)	-0.05 (0.06)	-0.11 (0.09)
Constant	0.55*** (0.21)	0.62*** (0.14)	0.55*** (0.13)	0.79*** (0.22)
# of obs	131	176	176	92
adjusted R-sq	0.1135	0.0754	0.1147	0.2694
sample	higher-than-ave or lower-than-ave	all	all	higher-than-ave

Notes: Results of ORIV (linear probability model) estimations with MAXBELIEFS treatment being the base group. All regressions include individual controls (risk-attitude, overconfidence, and IQ measure). ***, ** and * indicates significance at the 1%, 5% and 10% level, respectively. In Reg. (1) and (5) we focus on subjects who report own fines to be different from average fine. In Reg (4) and (8) we look at only those subjects who reported own fines to be higher than average fines.

lying behavior as the dice-in-the-cup paradigm commonly used in the lying literature, with the added benefit of capturing rich individual-level data. Second, we propose a simple and intuitive way of eliciting subjects' beliefs about own and average fine a violator might incur if caught lying. This difference is shown to be correlated with subjects' behavior in the partial information treatments.

From a substantive point of view, we find that communicating partial information about a fine distribution by using the minimum fine is the most effective at deterring crime among all considered schemes, which include another partial information scheme (the one in which the maximum fine is communicated) as well as several full information schemes. This result is robust to a replication and different types of analyses which focus on different deterrence objectives (e.g. reducing overall lying vs reducing the number of persistent liars). One of the reasons the minimum fine scheme is most effective at deterring lying comes from the

fact that announcing the minimum punishment removes conceivable optimistic expectations that the transgression might be particularly profitable even if one is caught lying. As a comparison, such optimistic beliefs are not ruled out when the maximum punishment is announced. The importance of this channel suggests that for the minimum fine scheme to be effective, the minimum punishment cannot be too low.

A natural next step would be to see whether this result replicates in a field study where partial information about fines for "real" crimes in the frame of a minimum is also most effective at deterring such behavior. As is the goal of controlled experiments, our study provides a clear recommendation for which treatments to test in the field (as well as what theoretical models predict should occur) such that policy relevant interventions are implemented in a productive way.

Furthermore, elicitation of subjects' beliefs reveals the mechanism behind our behavioral result. Subjects' tendency to lie is significantly and negatively correlated with their beliefs about their own fine, and the minimum frame induces higher beliefs about one's own fine as compared with the maximum frame. This result would not have been detected if one elicited only the average fines in the two partial information treatments, as is often the case, given that those have the opposite ranking: beliefs about the average fine in the maximum treatment are higher than those in the minimum treatment. We hope that this is also informative for policy makers, who may wish to better understand the belief structure of the constituents whose behavior they wish to influence.

We hope that our results will inspire more research on tools that emerge from advances in decision theory and their implications for real life situations.

References

Alm, A., B. Jackson, and M. McKee (1992). "Estimating the Determinants of Taxpayer Compliance with Experimental Data." National Tax Journal, Vol. 45(1): 107-114.

Alm, A., I. Sanchez, and A. De Juan (1995). "Economic and NonEconomic Factors in Tax Compliance." Working Paper.

Al-Najjar, N.I. and L. De Castro (2014). "Parametric representation of preferences," Journal of Economic Theory, 150:642–667.

Bebchuk, L. A. and Kaplow, L. (1992). "Optimal Sanctions When Individuals are Im-

perfectly Informed About the Probability of Apprehension." NBER Working Paper No. 4079.

Becker, G. (1968). "Crime and Punishment: An Economic Approach." Journal of Political Economy Vol. 76, No. 2: 169-217.

Brandts, J. and G. Charness (2011). "The strategy versus the direct-response method: a first survey of experimental comparisons." Experimental Economics 14: 375-398.

Calford, E. and DeAngelo, G. (2020) "Ambiguity Enforcement." Working Paper.

Casagrande, A., Di Cagno, D., Pandimiglio, A., and Spallone, M. (2015) "The Effect of Competition on Tax Compliance. The Role of Audit Rules and Shame." Journal of Behavioral and Experimental Economics, vol. 59: 96-110.

Cerreia-Vioglio, S., F. Maccheroni, M. Marinacci, and L. Montrucchio (2013). "Ambiguity and robust statistics," Journal of Economic Theory, 148:974–1049.

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E. and Camerer, C. (2018). "Econographics." CESifo Working Paper No. 7202.

Charness, G., Gneezy, U. and Imas, A. (2013). "Experimental methods: Eliciting risk preferences." Journal of Economic Behavior and Organization, 87: 43-51.

Chen, D., Schonger, M., and Wichens, C. (2016). "oTree—An open-source platform for laboratory, online, and field experiments." Journal of Behavioral and Experimental Finance, 9: 88-97.

DeAngelo, G. and Charness, G. (2012). "Deterrence, expected cost, uncertainty and voting: Experimental evidence." Journal of Risk and Uncertainty, 44: 73-100.

Denti, T. and L. Pomatto (2020). "Model and Predictive Uncertainty: A Foundation for Smooth Ambiguity Preferences," working paper.

Dwenger, N., Kleven, H., Rasul, I., and Rincke, J. (2016) "Extrinsic vs Intrinsic Motivations for Tax Compliance. Evidence from a Randomized Field Experiment in Germany."

American Economic Journal: Applied Economics.

Engel, C. (2016) "Experimental Criminal Law. A Survey of Contributions from Law, Economics and Criminology." MPI Collective Goods Preprint, No. 2016/7.

Erat, S. and Gneezy, U. (2012) "White lies." Management Science, 58, 723–733.

Engel, C. and Nagin, D. (2015) "Who is Afraid of the Stick? Experimentally Testing the Deterrent Effect of Sanction Certainty." Review of Behavioral Economics, Vol. 2: 405-434.

Feess, E., Schildberg-Horisch, Schramm, M., and Wohlschlegel, A. (2015) "The Impact of Fine Size and Uncertainty on Punishment and Deterrence: Evidence from the Laboratory." Working paper.

Friesen, L. (2012) "Certainty of Punishment versus Severity of Punishment. An Experimental Investigation." Southern Economic Journal, vol. 79: 399-421.

Fischbacher, U. and Follmi-Heusi, F. (2013). "Lies in Disguise: An experimental study on cheating." Journal of the European Economic Association, Vol. 11, No. 3: 525-547.

Fishbane, A., Ouss, A., and Shah, A. (2020) "Behavioral nudges reduce failure to appear for court." Science. 10.1126/science.abb6591.

Gilboa, I. and Schmeidler, D. (1989). "Maxmim expected utility with non-unique prior." Journal of Mathematical Economics, Vol. 18, No. 2: 141-153.

Gillen, B., Snowberg, E. and Yariv, L. (2018). "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." Journal of Political Economy, 127, No. 4: 1826-1863.

Gneezy, U., Meier, S. and Rey-Biel, P. (2011). "When and Why Incentives (Don't) Work to Modify Behavior." Journal of Economic Perspectives, 25(4): 191–210.

Gneezy, U. and Potters, J. (1997). "An Experiment on Risk Taking and Evaluation Periods." The Quarterly Journal of Economics, Vol. 112, No. 2: 631-645.

Gneezy, U., Rockenbach, B., and Serra-Garcia, M. (2013) "Measuring lying aversion". Journal of Economic Behavior and Organization, Vol. 93: 293-300.

Halevy, Y. (2007). "Ellsberg Revisited: An Experimental Study." Econometrica, Vol. 75, No. 2: 503–536.

Heifetz, A., Meier M., and B. Schipper (2006) "Interactive Unawareness." Journal of Economic Theory, Vol. 130 (1): 78-94.

Kahneman, D. and Tversky, A. (1979). "Prospect Theory: An Analysis of Decision Under Risk." Econometrica, Vol. 47, No. 2: 263-292.

Karni, E. and M. Viero. (2013). "Reverse Bayesianism': A Choice-Based Theory of Growing Awareness," American Economic Review, Vol. 103: 2790-2810.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005). "A smooth model of decision making under ambiguity," Econometrica, Vol. 73: 1849–1892.

Klibanoff, P., S. Mukerji, K. Seo, and L. Stanca (2019) "Foundations of ambiguity models under symmetry," working paper.

Machina, M. and M. Siniscalchi. (2013) "Ambiguity and Ambiguity Aversion." to appear in The Handbook of the Economics of Risk and Uncertainty, edited by M. Machina and W. Viscusi.

Nagin, D. and Pogarsky, G. (2003) "An Experimental Investigation of Deterrence. Cheating, Self-Serving Bias, and Impulsivity." Criminology, vol. 41: 167-193.

Salmon, T. and A. Shniderman (2019). "Ambiguity in Criminal Punishment." Journal of Economic Behavior and Organization, 163: 361-376.

Seo, K. (2009) "Ambiguity and second-order belief," Econometrica, Vol. 77:1575–1605.

Slemrod, J. (2019). "Tax Compliance and Enforcement." Journal of Economic Literature, 57(4): 904-954.

Tergiman, C., and Villeval, M.C. (2019) "The Way People Lie in Markets." Working Paper.

Tversky, A. and Kahneman, D. (1974). "Judgment under Uncertainty: Heuristics and Biases." Science, 185:1124-1131.