ESSAYS IN BEHAVIORAL AND EXPERIMENTAL ECONOMICS

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2018

vorgelegt von
Katharina Laske M.Sc.
aus Ludwigsfelde

Referent: Professor Dr. Dirk Sliwka

Korreferent: Professor Dr. Matthias Heinz

Tag der Promotion: 29.05.2019

Danksagung

Mein allererster und wichtigster Dank geht an meinen Doktorvater Dirk Sliwka für seine außergewöhnliche Betreuung, die sich durch fachliche Exzellenz, große Unterstützung und ständige Verfügbarkeit ausgezeichnet hat. Sein kritischer Geist gepaart mit konstruktiver Kritik, sein genuines Interesse an Erkenntnis und das damit verbundene Selbstverständnis dieses immer wieder zu hinterfragen sowie seine integre, gradlinige Haltung haben mich in meinem akademischen Werdegang geprägt.

Äußerst dankbar bin ich Matthias Heinz für die Übernahme des Koreferats, der nicht nur im übertragenen Sinne stets eine offene Tür hatte und mir während meiner gesamten Promotionszeit immer hilfsbereit zur Seite stand. Den offenen Meinungsaustausch mit Matthias und seinen guten Rat, wenn ich ihn brauchte, habe ich sehr geschätzt.

Von zentraler Bedeutung für das Anfertigen dieser Arbeit war die große Unterstützung meiner Koautorin Marina Schröder, die mich von Beginn an mit viel positiver Energie ermutigt, gefördert und begleitet hat. Vieles von dem, was ich in meiner Promotion gelernt habe, habe ich von Marina gelernt. Ihr gilt mein ganz besonderer Dank.

Zudem möchte ich mich sehr bei meiner Koautorin und Freundin Silvia Saccardo bedanken; sie machte meine Zeit in San Diego unvergesslich. Ein herzliches Dankeschön geht auch an Uri Gneezy, der meinen Forschungsaufenthalt in San Diego erst ermöglichte und mich von Beginn an aufgeschlossen und kollegial im Team aufnahm.

Des Weiteren möchte ich mich bei meinen tollen jetzigen und ehemaligen Arbeitskollegen Jesper Armouti-Hansen, Jakob Alfitian, Sebastian Butschek, Lea Cassar, Razi Farukh, Anastasia Danilov, Mira Fischer, Christoph Große-Bölting, Patrick Kampkötter, Felix Kölle, Kathrin Manthei, Lea Petters und Timo Vogelsang für die sehr gute Zusammenarbeit und positive Arbeitsatmosphäre bedanken. Ein großer Dank geht natürlich auch an unsere studentischen Hilfskräfte allen voran Sarah Adelfang, Julia Bohne, Emanuel Castillo Ruiz, Tobias Danzeisen, Markus Hartmann, Christine Herter, Annalena Holterhoff, Jonathan Holzbach, Leonie, Kern, Tarek Khellaf, Andrew Kinder, Ruth Neeßen, Miriam Reetz, Eva-

Maria Scheurenbrandt, Julia Schmidt, Sophia Schneider, Carolin Wegner, Paula Wetter und Sandra Wüst für ihr Engagement und ihre große Einsatzbereitschaft.

Und weil Leben nicht nur aus Arbeit besteht, bin ich meinen Freunden außerordentlich dankbar, die meine Freizeit während der letzten Jahre bereichert haben. Ganz besonders erwähnen möchte ich an dieser Stelle Caro und Ramona, Anja und Jako sowie die „Minions" David, Johanna, Johannes, Marina und Sonja für viele schöne Abende und die gemeinsam verbrachte Zeit, in der ich mich erholen und Energie aufladen konnte.

Ein großer Dank gilt meinen Eltern und meinem Bruder Tobias, die mich auf meinem Weg durch das Studium und die Promotion begleitet und stets unterstützt haben.

Mein letzter und größter Dank geht an meinen Freund Cornelius. Für seine grenzenlose Unterstützung und seinen starken Rückhalt, durch den sich vieles sehr viel leichter angefühlt hat. Für sein Interesse, sich in meine Forschungsfragen reinzudenken und diese mit seinem unvoreingenommenen Blick kritisch zu hinterfragen. Für seinen Humor, mit dem er mich so häufig zum Lachen bringt und der das Leben so viel schöner macht! Ich bin sehr glücklich, dass es ihn gibt.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

People frequently break rules. Such rule-breaking behavior is manifold. Not only does it manifest itself in unethical behavior, such as the violation of legal or moral standards, but also in innovating behavior, when e.g., existing conventions and habits are overcome. Consequently, depending on the situation and environment, rule-breaking may be either strictly condemned or highly socially approved. This thesis deals with the investigation and identification of the behavioral mechanisms underlying these two distinct classes of rule-breaking behavior.

In today's economy, innovation is an essential determinant of an organization's competitiveness and economic success. Hence, generating creative ideas and finding unorthodox approaches to existing problems is becoming increasingly relevant and understanding how to facilitate this idea generation process is clearly important. We therefore study how creative performance can be fostered, in particular focusing on two factors – incentives and expertise – and their role in individual creative performance (chapter 2 and 3).

However, in many other situations breaking the rules is not associated with productivity enhancements and carries a negative connotation. Examples include tax evasion or lying to achieve a personal advantage. In these cases, rule violating behavior is considered detrimental, because individuals, organizations or society as a whole may be harmed. Since such unethical behavior is prevalent in many economic transactions, policy-makers undertake considerable efforts to minimize it. We contribute to these endeavors by investigating the effectiveness of two deterrent mechanisms in reducing unethical behavior. In particular, we systematically vary the size of the fine and the probability of punishment and study their effects on decision-making in the context of lying (chapter 4).

Neoclassical economic theory assumes individuals to behave rationally at all times, i.e. they know their preferences, are perfectly informed and maximize their personal benefit. With the help of insights from psychology, sociology and, more recently, neuroscience behavioral economics identifies systematic deviations from these behavioral assumptions of decision-making and incorporates those findings into economic models. The goal is to provide a more realistic foundation of human decision-making in order to increase the explanatory power of economic analyses.

Empirical investigations of human behavior have been at the heart of behavioral economics and typically involve the use of experiments. Experiments are one widely applied methodology within the social sciences and can be regarded as a major source of knowledge that complements other sources such as theory or field data (Falk and Heckmann 2009). Experiments can be roughly divided into three categories. Laboratory experiments, field experiments and experiments on Amazon's Mechanical Turk marketplace (Mturk) (for a more detailed taxonomy see Harrison and List 2004). Laboratory experiments allow for a tight control of the decision environment, a fact that is hard to obtain in natural occurring settings. This is important because together with the exogenous assignment to treatment and control conditions, it allows to draw inference about the causal relationship of interest. In recent years, experiments on the internet platform Mturk have become increasingly popular as a source for recruiting participants for experiments. Compared to laboratory experiments, experiments on Mturk have the disadvantage that control of the decision environment is less tight. However, the subject pool is typically more diverse and due to technical ease a high number of subjects can be recruited within a short time period. Field experiments have the great benefit that they occur in the natural environment of the participants where they do not know that they are part of an experiment. Therefore, the external validity is highest in field experiments. In all chapters of this thesis we use experiments as a common element. However, depending on the best fit for answering our research questions we apply different experimental approaches and use either laboratory experiments, field experiments or experiments on Mturk.

In Chapter 2 we are interested in the effect of piece-rate incentives on idea generation.[1] While recent economic research has made great advancements in understanding the effect of incentives on performance in routine tasks (see e.g., Prendergast 1999, Lazear 2000), we know surprisingly little about the effectiveness of incentives on performance in creative tasks. There are some indications that the assumptions traditionally made in economic theory regarding the effect of

[1] Chapter 2 is joint work with Marina Schröder and based on Laske and Schröder (2018).

incentives on performance may not hold for creative tasks. For example, one important difference between both task types is their quality requirement: in routine tasks, quality is typically defined by the usefulness of a product for its predefined purpose. For these tasks, firms try to design incentives such that they achieve a high number of output units that meet a certain usefulness threshold. Creative products, however, in addition to being useful must also be original. Thus, for creative tasks the quality dimension is multi-dimensional involving both a *usefulness* and an *originality* dimension. To study whether the effect of piece-rate incentives may play out differently for creative tasks compared to routine tasks, we developed a new real-effort task that allows us to objectively measure three different dimensions of creative performance separately. These dimensions are quantity, usefulness and originality of the generated ideas. Between treatments, we vary whether piece-rate incentives are implemented and whether these are weighted by a quality component which either rewards usefulness or originality. We compare the results to a baseline with a fixed payment.

We observe that piece-rate incentives per se - whether weighted or not - result in an increase in the number of good (useful *and* original) ideas compared to a fixed wage. This positive effect of incentives is due to an increase in the overall number of ideas and to an increase in the variance of the quality of ideas. The higher the variance, the more likely it is that an individual comes up with an extraordinarily good idea. However, adding a usefulness-weighting component to the incentive scheme (paying workers according to how useful an idea is) reduces the positive effects of piece-rate incentives, because it leads to inefficient distortions of effort provision. Workers then focus too much on usefulness at the cost of quantity and originality of their ideas. Adding an originality-weighting to the incentive scheme (paying workers only for ideas that no one came up with before) simply adds complexity to the incentive system without being associated with any benefit.

Our findings provide interesting insights for organizations seeking to incentivize creativity. They imply that the most efficient way to incentivize idea generation may simply be by paying per idea without any ‘judgement’ regarding the originality or the usefulness of these ideas - even though they desire employees pay attention to multiple dimensions.

Besides heterogeneity in incentives, another factor that may be associated with the creative outcome is the level of an individual’s expertise in the focal field. Some researchers argue that creativity needs knowledge as a source of ideas from which original products can then be generated (Amabile 1998, Cropley 2006). They conclude that people who are familiar with the focal field will be more successful in finding good solutions. A different view, however, argues that experts rely on accustomed habits and think in familiar patterns which may block divergent thinking and reduce average creative performance (Wiley 1998). In chapter 3 we consider the case of outsourcing a creative task in a marketing context and study what is the most efficient way in terms of cost-benefit calculations when a company is in search of a creative solution: hire few expensive experts or many less costly non-experts? Our argument is as follows: even if, on average, experts produce a more creative outcome, average outcome is not the right measure. Since only the best ideas - the positive outliers - matter in creative tasks, it might be better to have more solutions and hire many non-experts than few experts. We explore this hypothesis by conducting a field experiment in which we use a new real-effort task that allows us to derive an objective measure for the quality of ideas. The difference between the two treatments is that experts were obtained via the world's largest freelancing platform *upwork*, where independent professionals from all of the world offer their expertise to potential customers and that non-experts consist of members from the platform *Mturk*.

We observe that for a given budget, hiring many non-experts instead of a few experts is more efficient in our setting. Although experts put significantly more effort into the task, as measured by time, this effort does not translate into better performance. This finding is consistent with the literature on creativity suggesting that creative performance is likely a probabilistic function of quantity (Laske and Schröder 2018). Our results provide a rationale for why an increasing number of firms have chosen to utilize crowdsourcing for idea generation purposes.

In chapter 4, we switch our attention to investigating factors that help to reduce rule violations in situations in which individuals, organizations or society as a whole may be negatively affected.[2] Several work has explored ways to prevent unethical behavior focusing on the intrinsic costs to individuals' self-image that arise from behaving unethically (e.g., Mazar, Amir, and Ariely 2008, Gneezy, Saccardo, Serra-Garcia, and van Veldhuizen 2015). In this paper we take a different approach and study how individuals react to interventions that focus on the extrinsic costs of unethical behavior. The standard economic model of crime (Becker 1968) assumes that the decision whether to perform a crime is similar to a choice of a lottery. There are three possible outcomes: the decision maker can choose not to commit a crime. If she chooses to commit a crime, then either she is successful, or she is caught and punished. In this model, the decision maker calculates her expected utility, weighs the utility of each outcome and the associated probabilities, and commits the crime if the expected utility of doing it is higher than the utility of not performing

[2] Chapter 4 is joint work with Silvia Saccardo and based on Laske and Saccardo (2018).

the crime. In order to identify what kind of policy would deter unethical behavior effectively, we use a systematic approach based on Becker's model. We investigate a situation in which participants can lie at another participants' cost to achieve an economic advantage and vary the size (high or low) and probability (5%, 10%, 25% or 50%) of punishment to study how these two factors affect unethical behavior in everyday life.

In all our one-shot experiments, when individuals are presented with only one set of parameters and are asked to make a decision only once, we find that lying decreases with the size of the fine. However, individuals are insensitive to changes in detection probabilities. Why is that the case? In two additional experiments, we show that sensitivity to detection probabilities only emerges when individuals can directly compare different detection probabilities in a within-participant design, or when they experience the same probability level over time in a repeated setting.

Our findings have several interesting implications: they suggest that harsher fines are likely to be a more successful means of deterrence of small-scale unethical behavior than increasing the probability of detection. Sanctions that are based on increases in detection probability may work under two conditions: First, providing individuals with a reference point could potentially increase the sensitivity to probabilities because this enables them to compare different probability levels. For example, instead of simply announcing a given detection probability, informing people that their chance of being audited has increased may help them to incorporate detection likelihoods into their decision. Second, such deterrence policies are likely to be effective for reducing unethical behavior in situations in which individuals receive frequent feedback on the outcome and likelihood of being audited. One example may be fare evasion in public transportation. However, such policies may be less effective in deterring unethical behavior when detection probabilities can only be presented via a description and feedback is rare, such as small-scale tax evasion.

Chapter 2

Quality through Quantity - The Effects of Piece-Rate Incentives on Creative Performance

"*The best way to have a good idea is to have many ideas" (Linus Pauling)*

2.1 Introduction

Creativity is among the most important, yet least understood factors that influence economic success. As a reaction to constant technological change and fierce competition, organizations are forced to permanently generate creative ideas to drive innovation. Hence, it is not surprising that fostering creativity is consistently rated as a primary concern of global top managers (see e.g., The Conference Board 2012, 2013, 2014). The important question is what organizations can do to successfully foster creativity. Despite its importance, there is very limited research on this topic. One reason for the scarcity of research is that creativity is notoriously difficult to define and to quantify. In this paper, we define creativity from a functional problem solving perspective, where good creative ideas must be both *useful* for a predefined purpose and *original* (Cropley and Cropley 2008). In our approach usefulness refers to the extent to which an idea meets the functional requirements of a task. Originality refers to the statistical infrequencies of an idea. We introduce a novel experimental design, which involves clearly defined and quantifiable performance indicators for these dimensions and study the effect of piece-rate incentives on idea generation.

While recent economic research has made advancements in understanding the effect of incentives on performance in routine tasks (see e.g., Prendergast 1999, Lazear 2000, Laffont and Martimort 2009, Gneezy, Meier, and Rey-Biel 2011), we know surprisingly little about the effectiveness of incentives on performance in creative tasks. There are some indications that the assumptions traditionally made in economic theory regarding the effect of incentives on performance may not hold for creative tasks. Consequently, the effect of incentives may play out differently for creative tasks compared to routine tasks. For example, intrinsic motivation is often considered to play a special role for creative performance. It has been claimed that intrinsic motivation will be conducive to creative work, while extrinsic incentives such as performance pay may be detrimental (see e.g., Amabile 1996). Furthermore, the effort-performance relation seems to be different for creative tasks compared to simple routine tasks. Unlike most other desirable workplace behaviors, effort provided for creativity does not necessarily always translate into better performance (Amabile 1996, Erat and Gneezy 2016). Additionally, pressure induced by incentives may actually undermine creative performance (Ariely, Gneezy, Loewenstein, and Mazar 2009, Azoulay, Zivin, and Manso 2011, Gross 2016). Another specificity distinguishing creative tasks from many routine tasks is the lack of clear definitions and objective measures. In the absence of a clear definition and objective measures, individuals may do not know what is expected from them and thus may, even if they wanted to, not be able to react to incentives (Byron and Khazanchi 2012, Charness and Grieco forthcoming). Furthermore, the lack of objective measures may lead to (strategic) distortions in the evaluation of creative performance, which potentially further undermines the effectiveness of incentives (Bradler, Neckermann, and Warnke forthcoming, Balietti, Goldstone, and Helbing 2016, Petters and Schröder 2017). Finally and importantly, the quality requirements for creative tasks are different from those for routine tasks. In routine tasks, quality is typically defined by the usefulness of a product for its predefined purpose. For these tasks, firms typically try to design incentives such that they achieve a high number of output units that meet a certain usefulness threshold. In addition to being useful, creative products must also be original. Hence, for creative tasks the quality dimension is multi-dimensional involving both a *usefulness* and an *originality* dimension. Firms therefore should seek to design incentive schemes for fostering creativity in a way that employees generate a high number of outliers which are at the same time highly useful and original. In our approach to understanding the effect of piece-rate incentives on creative performance, we focus on this special multi-dimensional characteristic, which is likely to have an influence on how incentives affect creative performance.

According to multi-tasking theory (see e.g., Holmstrom and Milgrom 1991, Lazear 2000), incentives in multi-dimensional contexts may entail undesirable distortion effects. The theory predicts that with the introduction of performance related incentives, individuals will focus on the productivity dimensions that can be easily measured and may reduce their effort in the productivity dimensions, which are not measured and thus provide lower returns for the agent. Recent empirical work shows evidence for such distortion effects in contexts involving a quantity and a one-dimensional quality measure (see Kachelmeier Reichert, and Williamson 2008, Hossain and List 2012). To the best of our knowledge, we are the first to focus on the multi-dimensional characteristic of quality measures for creative tasks.

In this paper, we contribute to the research on creativity by studying the effect of piece-rate incentives on performance in an idea generation task where usefulness *and* originality of creative ideas can be measured separately and objectively. Previous experimental research on the effect of piece-rate incentives on creative performance has led to inconclusive results. While some experimental studies find no effect of piece-rate incentives on creative performance (Eckartz, Kirchkamp, and Schunk 2012, Erat and Gneezy 2016) other studies find a positive impact (Kachelmeier, Reichert, and Williamson 2008). In this paper, we aim to provide further insights by focusing on the special multi-dimensional characteristic of creative work. This allows us not only to assess if and when incentives work, but also to provide novel insights into the mechanisms through which incentives affect creative performance.

We introduce a novel experimental design, in which participants are asked to illustrate words with the help of a given set of materials. They are instructed to create as many illustrations as possible (quantity) that can be recognized by independent raters (usefulness), and that are statistically infrequent (originality). In the task, participants do not receive a list or any specifications of words to illustrate or how to use the provided materials. Hence, they have to come up with both the words they want to illustrate and a way of how to illustrate these words. Between treatments, we vary whether piece-rate incentives are introduced and whether these are weighted by the usefulness or originality of ideas. We compare the results to a baseline with a fixed payment. The advantage of this experimental design is that we can establish objective creativity measures in a domain where this has so far been challenging. Furthermore, the experimental approach allows us to isolate the behavioral effect of creativity-based compensation under otherwise identical conditions.

We find that piece-rate incentives have a positive and significant effect on the number of high quality (useful and original) creative ideas, if designed appropriately. We uncover two channels for these positive effects: first, piece-rate incentives lead to an increase in the number of ideas generated; second, piece-rate incentives lead to an increase in the variance of the quality of ideas. The higher the variance, the more likely it is that an individual comes up with an extraordinarily good idea. However, we show that piece-rate incentives that reward quantity *and* usefulness can lead to an inefficient distortion in effort provision where participants focus too much on usefulness at the cost of quantity and originality.

Overall, we show that the special multi-dimensional characteristic of creative tasks is crucial to understand the effect of incentives on creative performance. Our results suggest that piece-rate incentives can work to foster creativity, but that organizations seeking to incentivize creative performance can be better off avoiding 'judgmental' incentives where the level of reward depends on the usefulness of ideas.

2.2 Experimental Design and Procedure

2.2.1 The Task

We propose a novel experimental design that allows for an objective assessment of performance in three dimensions of creativity, i.e. quantity, usefulness and originality. In this experimental design, we ask participants to illustrate words using several simple materials. The set of materials provided for each participant consists of one string, two O-rings, four wooden sticks, and twelve colored glass pebbles (see Figure 2.1 left picture). Participants can use some or all of these materials to illustrate words (see Figure 2.2 for example illustrations). Participants do not receive a list or any specifications of words to illustrate or how to use the provided materials; hence, in this task, they have to come up with both the words they want to illustrate and a way of illustrating these words. They can illustrate as many words as they want within a period of 20 minutes. After finishing an illustration, participants are instructed to take a picture using a pre-installed camera and to type in the illustrated word.[1] The advantage of the task is that it allows us to objectively measure multiple dimensions of participants' creative performance.

[1] A detailed description of how lab participants took a picture of an illustration is given in Appendix B.

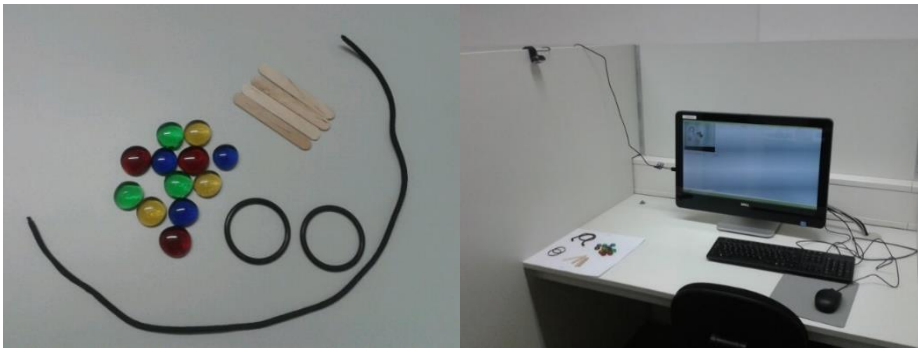


FIGURE 2.1: SET OF MATERIALS AND EXPERIMENTAL SETUP

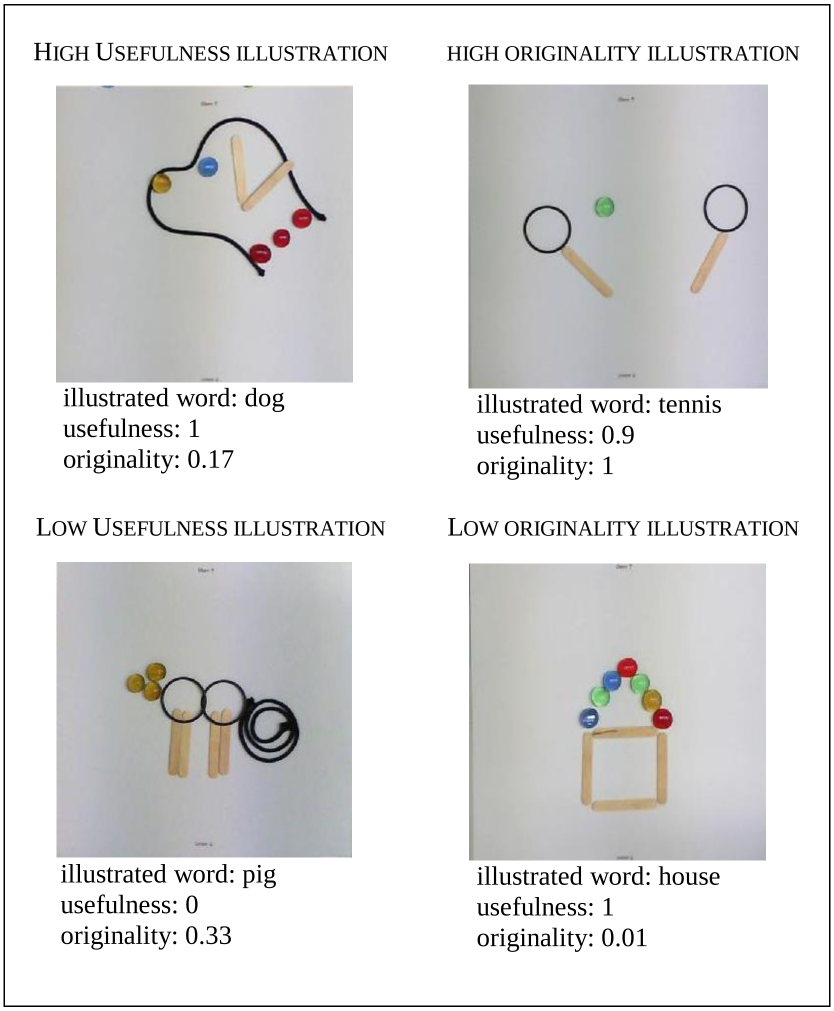
We measure *quantity* as the number of different words illustrated. That is, a participant scores high in this dimension if she illustrates a high number of different words. We directed participants to only illustrate single words (e.g., “tree”, “face”), to illustrate each word only once, and informed them that they are not allowed to use or illustrate any symbol found on the keyboard (e.g., “!”, “8”, “b”, “@”, “>” “+”). Illustrations of phrases consisting of more than one word (e.g., “tree in the woods”, “happy face”), multiple illustrations of the same word (e.g., two different illustrations of the word “house”), and illustrations including symbols from the keyboard (e.g., using “8” to illustrate the word “eight”) were not valid. We instructed participants about these rules and informed them that illustrations violating these rules would not be considered for payment. See Appendix B for the instructions.

To measure the *usefulness* of each valid illustration, we elicit the recognition rate by external raters through an online survey, which was conducted two weeks after the lab experiment. In this online survey, raters are provided with pictures of the illustrations from the lab experiment and are asked to type in the exact word that is illustrated.[2] We incentivize answers in this online survey by rewarding online raters €0.10 for each correct answer, which is defined as an exact match of a word illustrated by a lab participant and the answer by the online participant. See Appendix C for the instructions of the online survey and a screenshot of the online survey. Raters in the online survey did not take part in any previous related experiments and were blind to treatments. Each illustration was rated by at least ten online raters, and each rater rated a random sample of 50 illustrations.[3] For each illustration, we derive usefulness as the fraction of raters who correctly

[2] In the assessment of usefulness, we did not account for synonyms since we explicitly informed participants of both the lab experiment and the online survey in the instructions that only the exact match of the illustrated word by the lab participant and the answer by the online rater will be considered for payment. Spelling errors were not corrected. The special characters ä, ö, ü and ß were standardized to ae, oe, ue and ss, respectively. Capitalization of letters was not taken into account.

[3] We restricted the sample to 50 illustrations per rater to avoid overload of the raters.

identified the illustrated word. For instance, if 10 out of 10 raters recognize an illustrated word correctly, it would receive the highest usefulness score of 1. An illustration that is only identified by 1 out of 10 raters receives a usefulness score of 0.1. See Figure 2.2 (left column) for examples of high and low usefulness illustrations.



HIGH USEFULNESS ILLUSTRATION

illustrated word: dog
usefulness: 1
originality: 0.17

HIGH ORIGINALITY ILLUSTRATION

illustrated word: tennis
usefulness: 0.9
originality: 1

LOW USEFULNESS ILLUSTRATION

illustrated word: pig
usefulness: 0
originality: 0.33

LOW ORIGINALITY ILLUSTRATION

illustrated word: house
usefulness: 1
originality: 0.01

FIGURE 2.2: EXAMPLES OF ILLUSTRATIONS

We measure *originality* as the statistical infrequency of an illustrated word within the entire experiment. Specifically, we derive the originality of an illustration as the ratio of 1 and the number of times the same word is illustrated in the sample. For instance, a word that is illustrated once in the whole experiment, such as “tennis,” receives the highest originality score of 1. A word that is illustrated many times, such as “house”, which was illustrated 82 times, receives a low score of

0.012. See Figure 2.2 (right column) for examples of illustrations that scored high or low on originality.

Since only ideas that are at the same time highly useful and original can potentially result in innovation (economic implementation of an idea), we define the quality of an illustration as the product of usefulness and originality. In the later analysis, we focus on good ideas - those ideas where quality rates above or equal to the 75th percentile of all ideas with the baseline as reference group and on excellent ideas - those ideas where quality rates above or equal to the 90th percentile of all ideas in the baseline.

2.2.2 Treatments

Between treatments, we vary whether piece-rate incentives are implemented and whether these are weighted by a quality component which either rewards usefulness or originality. In the baseline treatment, all participants receive a €10 fixed payment, independent of their performance. After conducting the baseline treatment, we calibrated the size of the piece-rate incentives for the three treatments based on the performance in the baseline experiment. That is, given the performance in the baseline treatment, average payment would have been equal in all three treatments.

In the unweighted piece-rate treatment, participants are paid based on the number of words illustrated. For each illustration, they receive €0.60. In the usefulness-weighted piece-rate treatment, the piece-rate paid for an illustration depends on the number of raters who correctly identify the illustrated word. All illustrations are rated by 10 incentivized raters. For each illustration, participants in the usefulness-weighted piece-rate treatment receive €0.10 per rater who correctly identifies the illustrated word. Finally, in the originality-weighted piece-rate treatment, participants are paid based on the number of illustrations that are unique within a group of four participants.[4] For each illustration of a word that has not been illustrated by another participant in the randomly assigned group of four participants, participants receive €0.85.

Except for the description of the incentive scheme, all participants in our experiment receive the same information about the relevant dimensions of productivity and their measurement. Table 2.1 summarizes the treatments in our experiment and the number of participants in each of the treatments.[5]

[4] We used uniqueness within a group of four instead of originality as an incentive measure because of procedural reasons in running the experiment. However, this design element should not have an effect on subject's strategic considerations.

[5] For our analysis, we excluded one observation from the unweighted piece-rate treatment, because this participant only generated invalid illustrations.

TABLE 2.1: EXPERIMENTAL TREATMENTS

Treatment	Payment	Amount	N
Baseline	Fixed payment	€10	32
Unweighted piece-rate	Number of illustrations	€0.60 per illustration	31
Usefulness-weighted piece-rate	Number of raters who correctly identify each illustration	€0.10 per correct identification of each illustration per rater	30
Originality-weighted piece-rate	Number of unique illustrations (in a group of four)	€0.85 per illustration unique within a group of four participants	32

2.2.3 Procedural Details

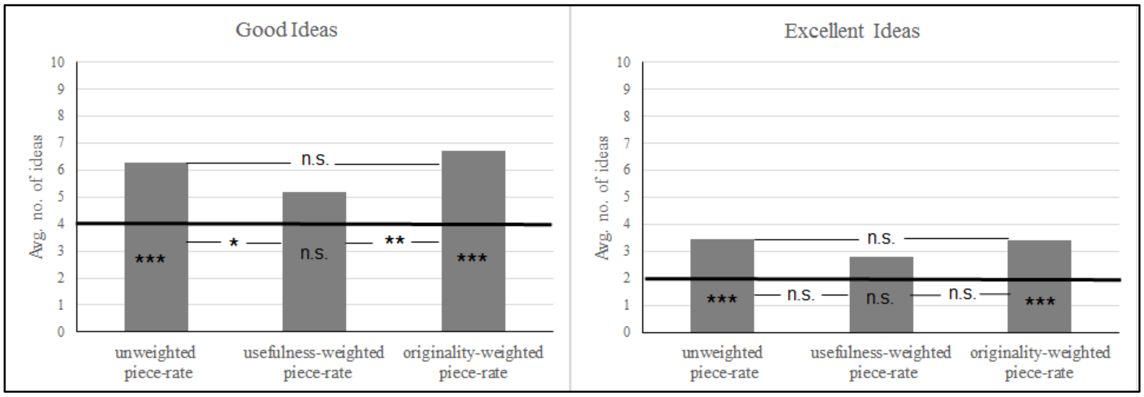
The experiment was conducted at the Cologne Laboratory for Economic Research at the University of Cologne. Participants were recruited with the online recruiting system ORSEE (Greiner 2004). We ran eight sessions in May 2014, with two sessions for each treatment condition. Participants were randomly seated in separated cubicles in the lab. To inform participants about the task, they received written instructions, which were read aloud by the experimenter. After the experimenter had answered all questions individually, the set of materials was handed to the participants. All illustrations of words had to be placed within a designated area on the desks. We told participants to place all materials that were not used for the illustration outside this area. Additionally, participants were instructed not to use any materials other than those provided by the experimenter. Once a participant made an illustration, she pressed a button on the screen of the computer so that the software would automatically take a picture of the designated area including the illustration. If participants were satisfied with the picture, they were asked to type in the word that they had illustrated and could then proceed with their next illustration. If they were not satisfied, participants could take another picture before proceeding. Figure 2.1 (right side) illustrates the cubicle in the laboratory, including the designated area in which participants provided illustrations and the web cam taking the pictures. As soon as the working time of 20 minutes was over, the experimental software automatically stopped and then initiated a questionnaire with some general demographic questions.

On average, each session lasted 40 minutes, and the average payoff was €14.43. The final payoff for each participant consisted of the money earned during the experiment and a standard show-up fee of €2.50. In all treatments, the money was paid out two weeks after the experiment, and participants could choose whether they preferred to collect the money in cash at the university or have it transferred directly to their bank account.

For the online survey, we recruited 540 participants from the same subject pool via ORSEE (Greiner 2004) and excluded participants who had previously taken part in the lab experiment. The online survey lasted about 20 minutes, and the average earnings were €4.50, including a €2.00 show-up fee. As in the lab experiment, participants had a choice between collecting the money in cash or a bank transfer.

2.3 Results

As a first step of our analysis, we study if piece-rate incentives work for creative tasks. To assess the success of the different incentive schemes, in Figure 2.3 we focus on the number of high quality ideas, where quality is defined as the product of usefulness and originality. By the term high quality ideas we subsume good and excellent ideas. Good ideas are classified by the average number of ideas above the 75th percentile of the product of usefulness and originality with the baseline as reference group. Excellent ideas are classified by the average number of ideas above the 90th percentile of the product of usefulness and originality with the baseline as the reference group for each of the treatments. The average number of good and excellent ideas from the baseline which does not involve piece-rate incentives is indicated by the solid black line.



Notes: Solid line indicates the average number of ideas in the baseline. Stars indicate the results from a two-tailed U-test *** p<0.01, ** p<0.05, * p≤0.1, n.s. not significant.

FIGURE 2.3: NUMBER OF HIGH QUALITY IDEAS BY TREATMENT

When we introduced piece-rate incentives in general (unweighted and weighted), participants reacted by producing a higher number of good ideas. This increase is only significantly for the unweighted- and originality-weighted piece-rate incentives (pairwise U-test, p<0.01[6]). We find very similar results, however on a lower level, when focusing on excellent ideas[7]

Result 1 *The introduction of piece-rate incentives increases the number of good and excellent creative ideas, but this positive effect is reduced if piece-rates are usefulness-weighted.*

Above we showed that piece-rate incentives can work, if designed appropriately. Next, we want to investigate through which channel incentives affect creative performance. As a first step, we focus on the average performance in the separate dimensions of creativity, i.e. on quantity, usefulness and originality. Table 2.2 reports the results of regression analysis exploring the effect of incentives on the three dimensions separately. Starting with quantity, columns 1 and 2 show the results on the overall number of illustrations per participant. In the model in column 1, we include a dummy for piece-rate incentives, which is equal to one whenever any form of piece-rate incentives (unweighted or weighted) is applied. Additionally, we control for whether a usefulness-weighting or an originality-weighting is implemented. The reference group is the baseline with fixed pay. We find that piece-rate incentives per se significantly increase the number of ideas generated. In particular, participants under a piece-rate scheme on average come up with about 9.5 ideas more compared to those receiving fixed pay. However, when a usefulness-weighting is added to the piece-rate incentive, this positive effect is strongly mitigated and participants on average submit only about 1.5 more ideas. This net effect is no longer different from zero (Wald test, *p*=0.42). Adding an originality-weighting to the incentive scheme does not have a significant effect on the number of ideas generated. Changes in performance in one dimension of creativity could be due to an incentive effect (increase in overall effort provided) or due to a distortion effect where individuals shift effort from one dimension to the other (see e.g., Holmstrom and Milgrom 1991). To control for such possible distortion effects, we add controls for productivity in the other dimensions of creativity in column 2 (i.e., controls for the average usefulness and the average originality of illustrations). Including these controls in the model, the coefficient for piece-rate incentives remains significant, but significantly reduces its size by 21 percent (χ^2-test, p<0.05). This suggests that introducing piece-rate incentives increases the quantity of ideas in large part

[6] We report two-sided p-values in the entire paper.

[7] To test whether our quality measure as the product of usefulness and originality actually captures the perceived quality of creative ideas, we also assess a subjective quality measure through an online-questionnaire. Our findings are robust to using this subjective measure in order to derive good and excellent ideas (see Appendix A).

through an incentive effect and to a smaller degree also through a distortion effect. The coefficient indicating a negative effect on quantity when a usefulness-weighting is implemented decreases by 32 percent as soon as we control for performance in the other dimensions (χ^2-test, $p<0.05$), but remains significantly negative. Hence, adding a usefulness-weighting to the incentive scheme leads to an additional distortion of effort but also reduces the incentive effect of piece-rate incentives.

TABLE 2.2: EFFECT OF PIECE-RATES ON SEPARATE DIMENSIONS OF CREATIVITY

Dependent variable:	Quantity		(Avg.) Usefulness		(Avg.) Originality	
	(1)	(2)	(3)	(4)	(5)	(6)
Piece-rate in general	9.467***	7.469***	-0.080**	-0.024	0.036	-0.018
	(2.427)	(2.289)	(0.038)	(0.022)	(0.040)	(0.025)
Piece-rate with usefulness-weighting	-8.031***	-5.423***	0.114***	0.030	-0.085**	-0.010
	(2.240)	(1.908)	(0.035)	(0.023)	(0.035)	(0.021)
Piece-rate with originality-weighting	-1.256	-0.419	0.038	0.020	-0.027	-0.005
	(2.471)	(2.263)	(0.033)	(0.019)	(0.034)	(0.022)
Female	-2.556	0.298	0.131***	0.064***	-0.090***	-0.022
	(1.771)	(1.738)	(0.026)	(0.016)	(0.025)	(0.015)
Quantity				-0.003***		0.002
				(0.001)		(0.001)
(Avg.) Usefulness		-30.887***				-0.541***
		(7.205)				(0.019)
(Avg.) Originality		-10.241		-0.608***		
		(11.753)		(0.019)		
Constant	18.042***	34.921***	0.440***	0.688***	0.298***	0.506***
	(1.693)	(6.042)	(0.027)	(0.025)	(0.031)	(0.037)
Observations	125	125	2,648	2,648	2,648	2,648
Clusters			125	125	125	125
R^2	0.193	0.316				

Notes: Columns (1) and (2) report OLS regressions with robust standard errors in parentheses. Columns (3)-(6) report random effects regression models with robust standard errors clustered on individual level. The dependent variable in columns (1) and (2) is a continuous variable indicating the overall number of illustrations produced, in columns (3) and (4) it is a continuous variable between 0 and 1 indicating the average usefulness of an illustration; in columns (5) and (6) it is a continuous variable between 0 and 1 indicating the average originality of an illustration. Piece-rate in general is a dummy variable coded as 1 whenever piece-rate incentives are introduced (unweighted or weighted). Piece-rates with usefulness-weighting and piece-rates with originality-weighting are equal to 1 in the corresponding treatments involving this weighting. Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. In columns (1) and (2) we control for average usefulness and average originality on individual level, in columns (3) and (4) we control for the overall number of illustrations and the average originality on illustration level and in columns (5) and (6) we control for the overall number of illustrations and the average usefulness on illustration level. Reference group is the baseline with fixed pay. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

Columns 3 and 4 display the results of a random effects regression model, with average usefulness as the dependent variable. In these models related to usefulness and in the following models related to originality (columns 5 and 6), we focus on results on illustration level. Column 3 reveals a significant decrease in average usefulness whenever piece-rate incentives are present. Putting this into perspective, on a scale from 0 to 1, the average usefulness per illustration drops by about 0.08 units. Given that the mean usefulness in the baseline is about 0.5, piece-rate incentives result in a 16 percent reduction in the average usefulness of ideas. Adding a usefulness-weighting component compensates for this negative effect of piece-rate incentives. Adding an originality-weighting does not have a significant effect on the average usefulness of ideas. As above, in column 4 we additionally control for performance in the other dimensions of creativity to test for the relevance of distortion effects. Adding these controls, we no longer observe any significant treatment effects. This suggests that the observed effects on usefulness seem to be driven by a distortion of effort rather than an increase in overall effort provision.

Columns 5 and 6 display the results of a random effects regression model with average originality of an illustration as the dependent variable. We find no significant effect of piece-rate incentives on originality. However, we observe that interacted with a usefulness-weighting of piece-rate incentives significantly decreases the likelihood that an idea is original by 8.5 percentage points. Surprisingly, an originality-weighting does not have any significant effect on the average originality of ideas. We conjecture that individuals may not be able to increase originality just by trying harder. Controlling for performance in the other dimensions of creativity (column 6) reveals that the negative effect of the usefulness-weighting seems to be driven by a distortion of effort.

Result 2 *Piece-rate incentives per se increase the number of ideas and decrease the average usefulness of ideas generated. Adding a usefulness-weighting results in a distortion of effort, which leads to an increase in average usefulness but a decrease in quantity and average originality of ideas. Adding an originality-weighting does not have a significant effect on the performance in the separate dimensions of creativity.*

Unlike in routine tasks where high quality is typically characterized by a high average usefulness and/or a low scrap rates, in creative idea generation tasks it is typically only the positive outliers that are relevant. Consider for example a company looking for a new slogan to advertise a new product. Only ideas that are positive outliers have a chance to lead to successful marketing, while ideas of average or low quality are irrelevant. To study the effect of incentives on the likelihood of finding such outliers, we focus on the within-subject standard deviations in the quality of ideas. In the baseline, the average within-subject standard deviation in the quality is equal to 0.056, while it is equal to 0.077, 0.075 and 0.073 in the unweighted piece-rate, usefulness-weighted piece-rate and originality-weighted piece-rate treatments, respectively. We find that the standard deviations in quality are significantly larger for all treatments involving piece-rate incentives compared to the baseline (pairwise U-test, $p<0.07$), but that there are no significant differences in the standard deviations in quality between the three treatments involving piece-rate (pairwise U-test, $p>0.50$).

Result 3 *Piece-rate incentives increase the variance of the quality of ideas an individual generates.*

Finally, we want to bring together the previous findings and test how relevant the performance in the separate dimensions of creativity and the variance in creative performance are to explain the observed effects of piece-rate incentives. Table 2.3 provides the results of an OLS regression with the number of good ideas in columns 1-3 and the number of excellent ideas in columns 4-7 as dependent variables.[8] Column 1 presents a model for the number of good ideas controlling for whether piece-rate incentives are present and whether a quality-weighting (either usefulness-weighting or originality-weighting) is implemented. In line with the non-parametric analysis above, we find that compared to the baseline with fixed wage, participants respond to piece-rate incentives by generating on average about 2.3 more good ideas. Adding a usefulness-weighting to the incentive scheme, however, significantly reduces this number to 1.1 more good ideas, while adding an originality-weighting does not have a significant effect on the generated number of good ideas. In the model in column 2, we include controls for the average performance in the separate dimensions of creativity and find that they are highly predictive for the number of good ideas. Consequently, the coefficient for piece-rate incentives, while remaining significant, substantially decreases by more than 60 percent when these controls are introduced (χ^2-test, $p<0.01$). When adding these controls, the coefficient for the usefulness-weighting is no longer significantly different from zero. Thus, the negative effect of adding a usefulness-weighting to a piece-rate incentive scheme seems to be due to the distortion effect that this weighting has on average performance in the separate dimensions of creativity. In the model in column 3 we control for the variance in quality. We find that this variance is highly predictive on the number of good ideas.

[8] The following findings are robust to some differences in the specification. First, the result is robust to changes in the threshold used to classify ideas as either good or excellent. We have conducted the same analysis with a threshold of 0.65, 0.70, 0.80, 0.85 and 0.95 and the results remain qualitatively the same. Second, the main results remain stable if we use a non-linear poisson model for the regression.

When adding this variable, the coefficient for piece-rate incentives is no longer significantly different from zero, suggesting that the positive effect of piece-rate incentives on the number of good ideas can be partly explained by an increase in the performance in the separate dimensions of creativity and partly explained by an increase in the variance of the quality of generated ideas.

TABLE 2.3: EFFECT OF PIECE-RATES ON NUMBER OF HIGH QUALITY ILLUSTRATIONS

Dependent variable:	No. of good ideas			No. of excellent ideas		
	(1)	(2)	(3)	(4)	(5)	(6)
Piece-rate in general	2.271***	0.905*	0.331	1.496***	0.898**	0.255
	(0.617)	(0.518)	(0.521)	(0.487)	(0.437)	(0.398)
Piece-rate with usefulness-weighting	-1.165*	0.00135	0.0250	-0.718	-0.0385	0.0217
	(0.668)	(0.541)	(0.524)	(0.553)	(0.476)	(0.430)
Piece-rate with originality-weighting	0.231	0.323	0.409	-0.0777	-0.0231	0.0884
	(0.738)	(0.547)	(0.516)	(0.576)	(0.482)	(0.425)
Female	-0.414	-0.577	-0.512	-0.276	-0.460	-0.334
	(0.478)	(0.396)	(0.363)	(0.383)	(0.354)	(0.285)
Quantity		0.215***	0.205***		0.126***	0.112***
		(0.0339)	(0.0285)		(0.0306)	(0.0228)
(Avg.) Usefulness		11.70***	6.962***		11.14***	5.654***
		(2.171)	(2.265)		(1.867)	(1.905)
(Avg.) Originality		8.957***	3.617		10.92***	5.409**
		(2.241)	(2.609)		(2.086)	(2.210)
Standard deviation in innovation			19.65***			21.99***
			(5.197)			(4.070)
Constant	4.100***	-7.706***	-4.839***	2.161***	-8.385***	-5.166***
	(0.477)	(1.842)	(1.772)	(0.388)	(1.652)	(1.493)
Observations	125	125	123	125	125	123
R^2	0.122	0.504	0.551	0.074	0.419	0.563

Notes: Robust standard errors in parentheses. The dependent variable in columns (1)-(3) is number of good ideas (quality ≥75th percentile) and in columns (4)-(6) it is number of excellent ideas (quality≥ 90th percentile). Piece-rate in general is a dummy variable coded as 1 whenever piece-rate incentives are introduced (unweighted or weighted). Piece-rates with usefulness-weighting and piece-rates with originality-weighting are equal to 1 in the corresponding treatments involving this weighting. Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. In columns (2), (3), (4) and (5) we control for the overall number of illustrations, average usefulness and average originality on individual level, Reference group is the baseline with fixed pay. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

The models in columns 4 to 6 provide results for the number of excellent ideas. Generally, these results mirror the results for good ideas (columns 1 to 3). Again, we observe that participants come up with about 1.5 more excellent ideas when they are confronted with piece-rate incentives. Adding a usefulness-weighting to the incentive scheme slightly decreases this positive effect but not significantly so. Adding an originality-weighting does not have a significant effect on the number of excellent ideas. As above, we control for the average performance in the separate dimensions of productivity in the model in column 5. Again, we find that average performance in the separate dimensions of creativity is highly predictive for the number of excellent ideas and that adding these controls substantially reduces the coefficient for piece-rate incentives by about 40 percent (χ^2-test, $p<0.1$). Again, we observe a residual positive effect of piece-rate incentives, which is explained by an increased variance in the quality of ideas (column 6).

Result 4 *Piece-rate incentives affect creative performance both through a change in performance in the separate dimensions of creativity (i.e. quantity, usefulness and originality) and due to an increase in the variance of the quality of ideas.*

2.4 Conclusion

The relevance of creativity as a driving force of economic growth raises the question of which factors influence creative performance. In this study, we examine the effect of piece-rate incentives on creative idea generation. We find that piece-rate incentives per se have a positive effect on the number of high quality ideas. This effect is due to an increase in the overall number of ideas and an increase in the variance of the quality of ideas generated.

Our results show that quality-weightings of piece-rate incentives are not advisable. Although organizations may be inclined to only pay for useful ideas instead of rewarding their mere number, such a weighting to piece-rate incentives can have a detrimental effect. We observe that adding a usefulness-weighting (i.e. paying workers a piece-rate that depends on how useful an idea is) mitigates the positive effect of piece-rate incentives. The negative effect of this weighting arises because a usefulness weighting leads to a distortion of effort. Participants focus on generating useful ideas, but this concentration on the usefulness entails a reduction in the overall number and in the average originality of ideas generated. Adding an originality-weighting to the piece-rate (i.e. paying workers only if they come up with ideas that no one had before) solely adds complexity to the incentive system, without bringing a benefit.

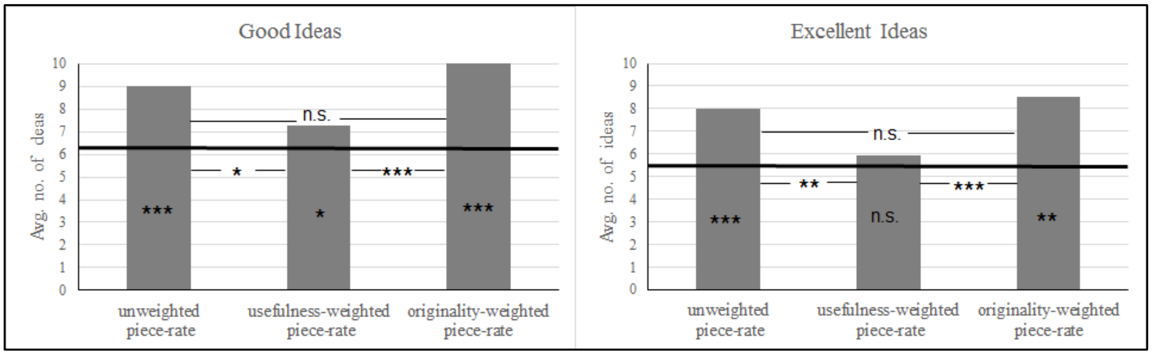
In this research project, we provide unique experimental evidence on the effect of incentives on quantity, usefulness and originality of creative ideas. Our results contribute to the understanding of how piece-rate incentives affect creative performance. We show that the special multi-dimensional characteristic of creativity – especially the relevance of simultaneous usefulness and originality of creative ideas – has a substantial impact on the effect of incentives on creativity. To the best of our knowledge, we are the first to reveal distortion effects between the separate quality dimensions of creative performance. Furthermore, we are the first to show that incentives have an impact on the variance in creative performance, which is associated with an increase in creative outliers. More research is needed to fully understand the behavioral mechanisms behind these effects.

Our research provides valuable insights for organizations seeking to incentivize creativity. We demonstrate that incentives can work to increase the number of high quality ideas generated. However, organizations should be very careful when designing such incentive schemes. Our results suggest that the most efficient way to incentivize idea generation may simply be by paying per idea without any ‘judgement’ regarding the originality or the usefulness of these ideas - even though they desire employees pay attention to multiple dimensions.

2.5 Appendix to Chapter 2

A Robustness Check

In addition to the objective measures of creativity we elicit a subjective measure for the quality of creative ideas. To obtain this measure, we asked two independent raters who were blind to the treatments to evaluate the creativity of all illustrations on an integer scale from 0 to 100. We derive the subjective quality of an illustration as the mean of these two evaluations. We conduct the same non-parametric analysis using this subjective measure as we did for the objective measure (see Figure 2.4). Using the score from this subjective quality assessments as the dependent variable does not change our results.



Notes: Solid line indicates the average number. of ideas in the baseline. Stars indicate the results from a two-tailed U-test *** p<0.01, ** p<0.05, * p<0.1, n.s. not significant.

FIGURE 2.4: NUMBER OF HIGH QUALITY IDEAS BY TREATMENT (SUBJECTIVE QUALITY MEASURE)

B Experimental Instructions for the Lab Experiment

Instructions

(Translation from German)

Welcome to this experiment!

Please carefully read the following instructions. If you have any questions, raise your hand. We will come to you and answer your question. Please do not **begin** the experiment until we ask you to do so. None of the other participants will receive information about your payoff. Communication with other participants is forbidden throughout the entire experiment. We also request that you switch off your mobile phone and remove it from the desk.

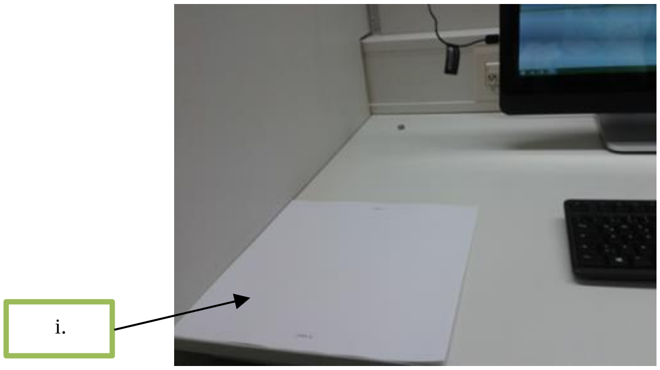
Task. - Immediately before the start of the task, you will receive various materials. The task consists of illustrating words with the provided set of materials. The goal is:

- To illustrate as many different words as possible,
- Which can be identified by others,
- And that the illustrated words are unique, meaning that they were not illustrated by any of the participants in the randomly selected four-person group.

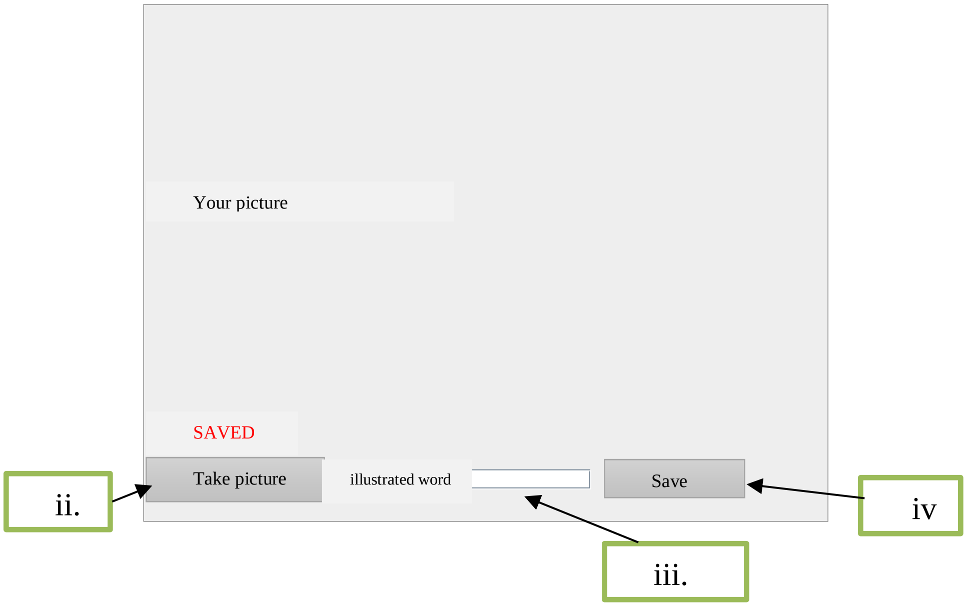
After the experiment, we will evaluate how well you achieved this goal.

Please proceed with the illustration of each word in the following manner:

i. Illustrate the word in the designated area using the provided materials.



ii. Take a picture of the illustrated word.
iii. Enter the word that you illustrated in the field “illustrated word”.
iv. Save the picture by clicking on the “save” button.



Please keep the following in mind:

- Use **only** the materials provided.
- **For each** illustrated word, you can use all of the materials or a selection of them.
- The illustration of the word should only be placed **within** the designated area on the sheet of paper (only this area will be captured by the camera).
- Make sure that your illustration is made in the correct **direction** (the sheet is marked “top” and “bottom”).
- Make sure that your **hands are not visible** in the designated area.
- Keep any **unused materials outside** of the designated area.
- Illustrate only **one** word at a time. This means that the name of the picture should only consist of **one word**. Terms that consist of multiple words are not permitted and will not be evaluated.
- You may only illustrate each word **once**.
- Your illustrations may not include any symbol that is depicted on the keyboard (for example, illustrations that include “→”, “8”, “b”, “@”, “>” or “+” are not permitted).

Time. - You have a total of 20 minutes for this task. After this time has expired, we ask you to answer the questionnaire before the end of the experiment.

Payment. - [This part is different with regard to the four treatments of the experiment]

Baseline: You are paid €10 for this task. In addition, you receive a show-up payment of €2.50. You will receive your payment two weeks after the experiment takes place. You can choose whether you would like to receive an electronic transfer or pick up the payment in cash.

Unweighted piece-rate: You are paid €0.60 for each admissible word that you illustrate. You also receive a show-up fee of €2.50. You can choose whether you would like to receive an electronic transfer or pick up the payment in cash.

Usefulness-weighted piece-rate: After this experiment, we will show the pictures of all of the admissible words you illustrated to other people. These other persons have not participated in this experiment or similar experiments. The task assigned to them is to identify the illustrated words using the pictures taken in the experiment. These other persons only receive a positive payout if they enter exactly the word that you saved along with the respective picture.

Each word will be presented to ten other people. We measure how many of these ten people correctly identify the respective word. For each illustrated word, you are paid **€0.10** for each person who correctly identifies it. That means you can earn up to **€1** for each illustrated word, assuming it is correctly identified by each of the ten people. In addition, you receive a show-up payment of €2.50. You will receive your payment two weeks after the experiment takes place. You can choose whether you would like to receive an electronic transfer or pick up the payment in cash.

Originality-weighted piece-rate: After this experiment, you will be randomly assigned to a group of four people who participated in the same experiment. For each admissible word that you alone in the group illustrated, you are paid €0.85. If at least one other person in the group illustrated the same word, then you receive €0 for illustrating this word. In addition, you receive a show-up payment of €2.50. You will receive your payment two weeks after the experiment takes place. You can choose whether you would like to receive an electronic transfer or pick up the payment in cash.

C Instructions for Online Survey to Assess Quality

Instructions

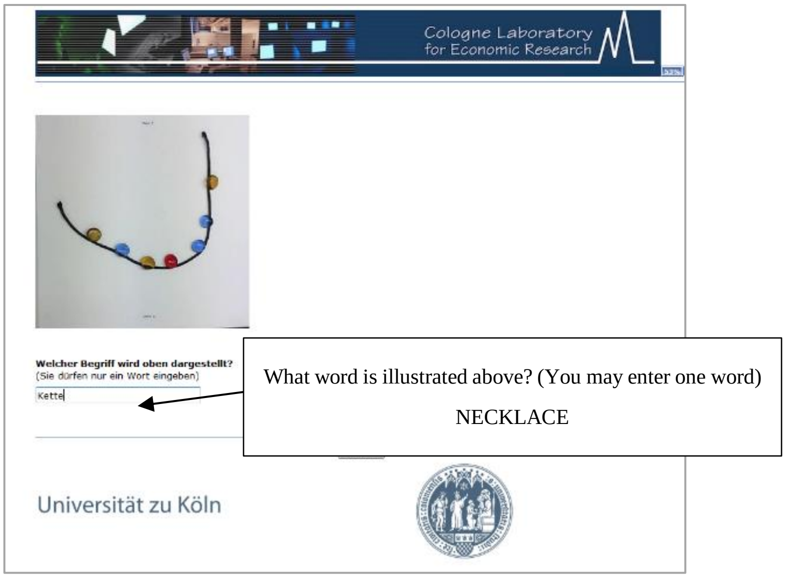
(Translation from German)

Please carefully read the following instructions. If you have any questions about these instructions or if you have any trouble with the experiment, please contact us by e-mail at internetexperimente@wiso.uni-koeln.de. Please note that you are not allowed to go back to a previous page at any time during the experiment. Next, you will see 50 consecutive pictures on your screen. These pictures were taken by participants in a prior experiment. These participants' task was to illustrate words using the materials provided. The words could be chosen freely and had to consist of only one word.

Your Task. - **Your task is to identify the illustrated words.** In order to receive payment for a picture, you must enter the **exact** word that the other participant assigned to that picture. If you do not make an entry for a picture, or if the word you enter does not exactly correspond to the respective word assigned by the other participant, then you do not receive any payment for this picture. Please take note of the fact that each of the illustrated terms consists of only **one word**. Your entries may also only consist of one word each. If you enter more than one word for a picture, it will be classified as 'not identified.'

Please also note that the words were illustrated by different participants. This means that it is possible to see more than one illustration of the same term.

Payment. - You will receive your payment only if you complete the entire experiment. You receive €2.00 for participating in the experiment. In addition, you receive €0.10 for each picture that you correctly identify. At the end of the experiment, you can choose whether you would like to receive an electronic transfer or pick up the payment in cash.



Cologne Laboratory for Economic Research

Welcher Begriff wird oben dargestellt?
(Sie dürfen nur ein Wort eingeben)

Kette

Universität zu Köln

What word is illustrated above? (You may enter one word)

NECKLACE

FIGURE 2.5: SCREEN OF QUESTIONNAIRE TO ASSESS QUALITY

Chapter 3

Creative Solutions: Expertise versus Crowd Sourcing

3.1 Introduction

Imagine you want to publish a book and need to find a title. You have a given budget to hire help with finding a good title. With this budget, you can either hire one expert (e.g., a copywriter) in the field, or a few non-experts. Your only interest is to have the best possible title for your book. Who is more likely to produce this: the expert or the non-experts?

Firms spend enormous amounts of money on R&D in which new technologies, products and services are created.[1] For many of these new developments, expertise and deep understanding of science and technology is a necessary condition. However, companies also expend large sums of money in areas such as marketing (like the title example), to acquire creative expertise from external providers such as advertising agencies or consultants. For instance, the ten world's biggest advertising agencies had a combined gross income of more than $25 billion in in 2017.[2] Such creative domains require little technical knowledge. Would experts produce a better creative output than non-experts in such situations, and relatedly, would employing expensive experts be efficient in terms of cost-benefit calculations?

Two possible effects of expertise on creativity may play a role. On the one hand, creativity may need knowledge as a source of ideas that can be used to generate novel products (Cropley 2006).

[1] The top 10 countries R&D spending was approximately 1.7 trillion dollars in 2018 (IRI, Statista, 2018).

[2] https://co.agencyspotter.com/50-largest-marketing-companies-in-the-world/

Amabile (1998) notes that domain-specific expertise is the “raw material” for creative ideas and people who are familiar with the focal field will be more successful in finding good solutions. On the other hand, however, experts may rely on accustomed habits and think in familiar patterns. Proponents of this view argue, that experts are more likely to be bound to the current thinking in their field, which blocks divergent thinking and reduce the potential for generating creative ideas (Wiley 1998). Thus, it is unclear whether experts will produce higher quality creative output than non-experts and whether employing experts in a domain such as marketing is efficient in terms of cost-benefit calculations. Even if, on average, experts produce a more creative outcome, average outcome is not the right measure. In many creative tasks, only the best ideas—the positive outliers—matter. Thus, having more solutions might be better, even if the average quality is lower.

We conduct a field experiment using a new real-effort task that allows us to derive an objective measure for the quality of ideas in a marketing context. Experts consist of members from the platform upwork and non-experts from MTurk. The platform upwork is the largest freelancer platform globally with specializations such as marketing, website designing or translation services, with about 12 million freelancers offering their expertise to potential customers. In contrast, MTurk is a crowdsourcing webpage by Amazon’ Mechanical Turk without any specialization opportunities that allow companies and researchers to post tasks that require a human to accomplish including writing contents for websites, researching data details or transcribing audio recordings.

The task for all hired workers was to come up with a creative title for a given short video clip (70 seconds). To measure quality, we recruited students from the lab. We showed each participant, four randomly selected titles from those created by the experts and non-experts and asked them to choose one title and watch the corresponding video afterwards. Thus, participants take real decisions involving opportunity cost of time. This design allows us to quantify the quality of a creative title by measuring the click rate.

Our key result is that experts invest significantly more effort, as measured by time, but that this endeavor does not translate into better performance. For example, seven out of the best ten ideas in our experiment were generated by non-experts. Taking into account that non-expert are cheaper to recruit, we observe that in our setting it is more efficient to hire few non-experts to generate the best creative solutions, rather than the more expensive experts. Our finding is in line with prior research that quantity breeds quality - meaning that the more ideas are generated, the higher the probability that some of them will be very creative (e.g., Laske and Schröder 2018).

Our paper is related to the growing literature in economics about creativity and incentives, i.e. the comparison of different incentive schemes (Eckartz, Kirchkamp, and Schunk 2012, Erat and Gneezy 2016, Bradler, Neckermann, and Warnke forthcoming), the dimensions that should be incentivized (Kachelmeier, Reichert, and Williamson 2008, Laske and Schröder 2018), how incentives interact with the type of the ideation task (Charness and Grieco forthcoming), how the magnitude of the incentive effects the creative outcome (Ariely, Gneezy, Loewenstein, and Mazar 2009) or how the time horizon of the incentive - short-term vs. long-term - influences individuals ideation performance (Azoulay, Zivin, and Manso 2011, Ederer and Manso 2013). Two studies are most relevant to our work, studying the role of expertise in a context where profound knowledge is indispensable. Jespersen and Lakhani (2010) examine successful solution in science problem-solving contests, finding that the provision of a winning solution is positively related to increasing distance between the solver's field of technical expertise and the focal field of the problem. Franke, Poetz and Schreier (2014) experimentally study whether individuals with expertise in the focal area provide better ideas for new product development compared to individuals with expertise in a different but similar market. They observe that the latter come up with product ideas with lower potential for immediate use, however, they demonstrate substantially higher levels of novelty. Similar to Jespersen and Lakhani (2010), the authors find that this effect is particularly pronounced when the distance between the two similar markets increases.

Building on this literature, we study the efficiency of experts in the creative domain. Our results provide a rational for why an increasing number of firms have chosen to utilize crowdsourcing for idea generation purposes. Forbes, for examples, states on its website that 85% of the 2014 Best Global Brands have used crowdsourcing in the last ten years[3].The underlying concept is to outsource the ideation phase to a large population of non-professionals in order to use the "wisdom" of the crowd.

3.2 Experimental Design and Procedure

3.2.1 Procedural Details

In October 2015, we recruited 30 experts on the platform upwork. In order to register on upwork.com as an expert, people need to create a profile which is thoroughly reviewed by the

[3] https://www.forbes.com/sites/steveolenski/2015/12/04/the-state-of-crowdsourcing/#2071416c55ee

platform. Once the information and professional credentials are verified (e.g., by entering codes for test results such as the Cambridge Certificate) the profile gets approved. People can take different test such as PHP Test, English Spelling Test or HTML. Upwork is the world's largest freelancing website, where independent professionals from all over the world can offer their expertise to potential customers. Upwork has 12 million registered freelancers providing companies with over 3,500 different skills.[45] By serving five million clients each year, freelancers are earning more than $1 billion.[6] The advantage of such platforms is that workers enjoy freedom and flexibility by working remotely, while business or private customers can benefit from the access to a large talent pool. As non-experts, we recruited 90 workers on Amazon's Mechanical Turk marketplace (Mturk), which became a standard source of participants for experiments. Previous studies show that the findings of studies run by Mturk are similar to the result in a more standard lab or field setting (Horton, Rand, Zeckhauser 2011, Amir, Rand, and Gal 2012, Goodman, Cryder, and Cheema forthcoming).

In treatment expert, after we had created the job post, we sent selected experts a message to their upwork account in which we invited them to apply for the job post. In that message we informed workers that the job will take less than an hour and has to get done within 24 hours after receiving the briefing. We also informed freelancers that they will be compensated with a fixed amount of $60 plus bonus opportunities for doing this job. For those who sent us a proposal, we send them a personalized link to the task.

Similarly, we posted the job on Mturk to recruit non-experts and announced it the same way as we did on upwork, with a description stating that participants would earn $1.5 plus bonus opportunities. Except of the payment, the instructions were identical between the treatments (see Appendix A for the instructions). As soon as workers had submitted a title, they were asked to fill out a questionnaire with some demographic questions.

The money was paid out within two days after workers had submitted a title. For the quality elicitation of the titles, we recruited 600 student raters via ORSEE (Greiner, 2004). The quality elicitation lasted about 3 minutes and student raters received €2 fixed pay.

[4] https://www.techlist.pk/pakistan-upwork-social-impact-program/

[5] https://www.upwork.com/about/

[6] https://www.upwork.com/about/

3.2.2 Treatments

The experts in our study are freelancers on the online platform upwork with a specialization in copywriting. To restrict the definition even further, we only recruited freelancers with an hourly wage of at least $60, assuming that the hourly wage signals the degree of expertise in the specialized area.[7] To minimize heterogeneity between the two treatment samples we only recruited US residents. In treatment *expert*, workers received a fixed amount of $60 per created title, independent of their performance.

In the corresponding treatment *non-expert*, participants were Mturk workers with US residency. All non-experts received a $1.5 fixed payment.[8]

To incentivize both experts and non-experts to do their best, we offered a $400 bonus for the worker who comes up with the best title. Table 3.1 summarizes the treatments in our experiment and the number of participants in each treatment.

TABLE 3.1: EXPERIMENTAL TREATMENTS

Treatment	Recruitment pool	Payoff	N
expert	Freelancers from the online platform upwork.com specialized in copywriting with an hourly wage of at least $60 with US residency	$60	30
non-expert	Mturkers with US residency	$1.50	90

3.2.3 The Task

We use a novel task to quantify the quality of creative performance. We asked participants to come up with a creative title for a 70-second video clip[9] on various forms of cheating. We informed participants that we would collect titles from different people and that the creator of the best title would be additionally rewarded a bonus of $400. Participants were informed that the quality of the title would be evaluated according to the number of clicks the title would generate among laboratory participants. See Appendix A for the instructions. For the quality measure, we recruited 600 raters via ORSEE (Greiner 2004). The student raters were blind to the source of the

[7] For each job category, upwork provides price ranges for hiring freelancers on entry, intermediate and expert level. For hiring people on expert level in the job category *copywriting* the platform suggests a payment of more than $46.50/hour.

[8] The size of the fixed payment for Mturkers was calibrated to data published by Horton and Chilton (2010), according to which fifty per cent of workers are willing to perform tasks for approximately $1.38 per hour. Since we only recruited US located Mturkers we increased this amount to $1.5.

[9] see https://www.youtube.com/watch?v=GwgtwY3oL4g&feature=youtu.be for the video

title ideas (experts vs. non-experts) and did not take part in any previous related experiments. We provided each rater with four randomly drawn titles and asked them to first click on the title of the video they wanted to watch and then watch the video. Thus, raters made real decisions involving opportunity cost of time.

On average, each title was seen by 20 raters, each time in a different combination with three other randomly selected titles. We derive the quality of each title as the fraction of raters who clicked on it such that quality ranges from 0 to 1. For example, if a title was clicked by 8 out of 20 raters, it received a quality score of 0.4. The entire list of titles and their score can be found at Appendix B. See Table 3.2 for examples of the three best and the three worst created titles in our experiment.

TABLE 3.2: EXAMPLES OF THE THREE BEST AND WORST TITLES

Best three titles	**Quality score**	**Worst three titles**	**Quality score**
1. Liars, Cheats & Thieves: The Truth about Human Character	0.61	1. ALL KINDS OF CHEATERS	0.05
2. Immoral Statistics: Breaking Society's Rules	0.60	2. We know you cheat!	0.05
3. Cheating, the new social norm?	0.55	3. Video infographic on cheating	0.05

3.3 Results

Do experts perform better than non-experts in a task for which prior knowledge is not necessarily required and only the best entries are important? As explained above, we define *quality* by the fraction of raters who clicked on a particular title in order to watch the related video. By design, the average overall quality was 0.25. The minimum quality in our experiment was 0.05 and the maximum quality was 0.61. We observe slightly higher average quality in *expert* (0.268) than in *non-expert* (0.244) titles.[10] This difference, however, is not statistically significant at any conventional level (U-Test, *p*=0.28; we report two-sided p-values in the entire paper). Figure 3.1 illustrates the results, with the vertical lines representing the means.

[10] Descriptive statistics on socioeconomic background data for experts and non-experts can be found in Appendix C. The share of woman, the mean age, the share of white people, and the average education level are significantly higher in the expert group. As expected, the self-rated expertise level of experts is higher than that of the non-experts.

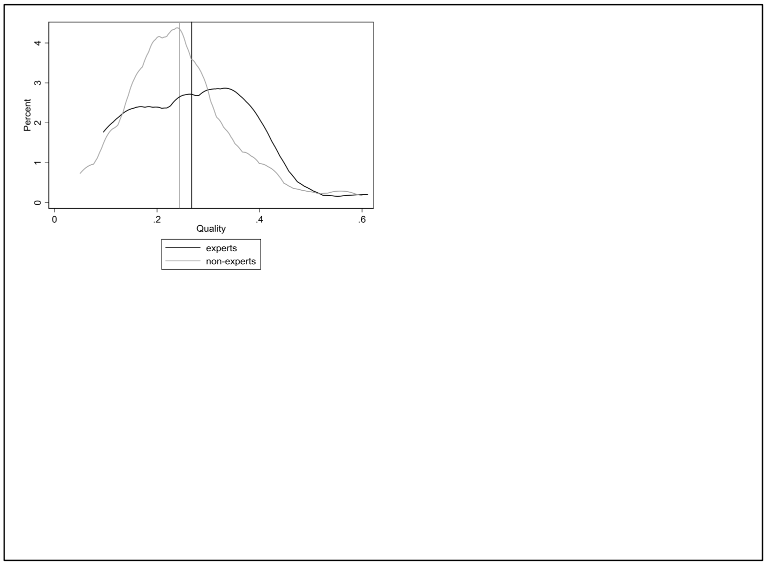


FIGURE 3.1: CLICK RATES BY TREATMENT

Table 3.3 reports the results of an OLS regression analysis with average quality in columns (1) and (2) as dependent variables. In line with the non-parametric analysis, we observe a slight but insignificant increase in performance of experts.

Interestingly, although experts do not seem to perform better on average, they do spend much more time on the task. On average, non-experts took 4.5 minutes to complete the task, whereas experts took 105 minutes. This difference is highly significant (pairwise U-test, $p<0.001$). Because this result regarding average time is driven in part by outliers, we conduct a median regression analysis to remove this outlier bias. Columns (3) and (4) of Table 3.3 display the results of a median regression analysis with submission time in minutes as the dependent variable. In line with the results from the non-parametric analysis column (3) reveals a significant increase in the median submission time whenever an expert created a title. Putting this finding into perspective, on a scale from 0 to 1380, the median submission time for a title increases by about 14 minutes when an expert versus a non-expert created a title. Adding sociodemographic controls in column (4) qualitatively does not alter the results. We conclude that experts try much harder to come up with a creative title, as measured by time, but that this endeavor does not translate into better performance. This finding is consistent with Amabile's (1996) premise that creativity does not emerge from simply trying harder. The platform upwork provides proxies for a worker's expertise. In addition to workers' self-reported field of expertise and hourly wage, we also collected information on the number of projects completed and hours worked. However, we observe no significant differences for any of these additional measures (i.e., they were not a more powerful predictor of performance).

Table 3.3: Regression analysis of expertise on Quality, Time and Productivity

	Average quality (in percent)		Average submission time (in min)		Average productivity per time unit	
	(1)	(2)	(3)	(4)	(5)	(6)
Expert	0.024	0.020	13.950**	14.287**	-0.060***	-0.063***
	(0.025)	(0.034)	(6.138)	(6.751)	(0.011)	(0.016)
Female		-0.005		0.195		-0.011
		(0.020)		(0.913)		(0.011)
Age		0.000		0.027		-0.001
		(0.001)		(0.085)		(0.001)
Expertise (self-rated)		0.006		-0.513		0.006
		(0.010)		(0.325)		(0.005)
Education		yes		yes		yes
Ethnicity		yes		yes		yes
Constant	0.244***	0.202***	2.800***	3.760	0.092***	0.087***
	(0.011)	(0.055)	(0.392)	(2.931)	(0.007)	(0.025)
Observations	120	119	120	119	120	119
R^2	0.009	0.080			0.157	0.243

Notes: Estimates in col. (1-2) and (5-6) are based on OLS regression. Estimates in col. (3-4) are based on LAD regression. Robust standard errors in parentheses Dependent variables: col. (1 and2): title quality in percent; col. (3 and 4) time between reading the task and submitting a title in minutes; col. (5 and 6): productivity per worker. The dummy Expert is equal to 1 if a worker was recruited on the freelancer platform upwork.com and 0 if a worker was recruited on Mturk. Self-rated expertise is a variable ranging from 1 (very low) to 5 (very high). Education is a variable indicating the highest degree of education differentiating between high school diploma, bachelor's degree, master's degree or higher and other. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

In a competitive economic environment, such as the advertising market, companies must generate creative ideas while not taking too long for one idea in order to be able to process the next task. Therefore, we now focus on a joint analysis of quality and submission time. That is, we consider the quotient of quality and submission time to receive a measure for productivity ($\text{productivity} = \frac{\text{quality}}{\text{submission time}}$). We find that productivity is significantly higher in the *non*-expert compared to the expert treatment (pairwise U-test, *p*=0.000). This result is confirmed by columns (5) and (6) of the OLS regression analysis in Table 3.3.[11] In particular, experts on average are 0.061 units less productive compared to non-experts. Given, that the mean productivity is

[11] The significance level remains unchained if we conduct the regression analysis without the outliers. Table upon request.

0.093 this means a productivity decrease of 66 percent. Adding sociodemographic controls does not affect the main qualitative results.

Next, we investigate whether it is more efficient to hire few experts or many non-experts to work on a creative task. To that aim we take into account the different amounts of fixed payment and calculate the cost per click as another productivity measure. While experts earned $60, non-experts earned $1.5 each without considering the bonus. The cost per click for a title created by an expert was, on average, $11.39, whereas the cost per click for a title created by a non-expert was $0.31. This difference is highly significant (pairwise U-test, *p*=0.000). To account for the fact, that this result depends on the parametrization of the payments we calculate the break-even point at which it gets more efficient to hire experts. If non-experts costed more than $55.95 it would be more efficient to hire experts in our experiment.

As discussed above, only the best creative ideas are relevant for the innovational process. We find that out of the best ten titles from this experiment, only three were generated by experts. For the sake of our analysis, we consider ideas with quality above the 90th percentile as excellent. We find the likelihood that an expert creates an excellent creative idea is equal to 0.132 and that of a non-expert is equal to 0.089. This difference is not significant (*p*=0.64).

3.4 Discussion and Conclusion

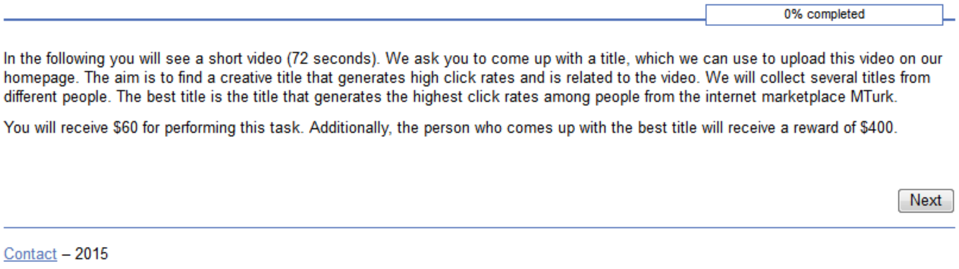
We study the efficiency of expertise in finding creative solutions in an area in which prior knowledge is not necessary. In our setting, we observe that for a given budget, hiring many non-experts instead of a few experts is more efficient. We also find that although experts put significantly more effort into the task, as measured by time, this effort does not translate into better performance. This finding is consistent with the literature on creativity suggesting that creative performance is likely a probabilistic function of quantity (Simonton 2003; Laske and Schröder 2018).

Our results are in line with the trend of using the creative potential of the population to fuel new product development, adopted by many firms. One prominent example is McDonalds, which instead of contracting food or industry experts in 2014, invited the public to submit ideas for the types of burgers they wanted to be offered in store. People could create their individual burgers online and the rest of the country could vote for the best ones, which were ultimately sold in the branches. Similarly, Threadless, a t-shirt company based in Chicago, has relied on the creative potential of the crowd since it was founded in 2000. Instead of employing professional designers, the company has asked people to submit designs via their crowdsourcing platform. Once an idea is posted, people start voting for it and leaving comments. Based on the average score and user feedback, about ten designs are selected each week, printed on clothes and other products and sold worldwide.

We acknowledge two important limitations to our study. In this experiment we had to rely on self-reported measures such as the domain of expertise and hourly wage to recruit experts from the platform upwork. Bradler (2016), however, observes in her study substantial misjudgments in agent's self-assessment for creative performance such that many agents held wrong beliefs about their relative ability in creative tasks. In line, we observe in our data only a small positive and insignificant correlation between self-rated level of creativity and the quality of the submitted ideas (ρ=0.117). Future research could expand the current study by recruiting employees from a real marketing agency in which employees have already proven their creative thinking skills and face real-world incentives such as reputation or promotion concerns. Moreover, a decisive success factor is the company's ability to identify the top ideas among all submitted ideas. The process used in our experiment might not be applicable to companies in all situations as it requires a representative sample of its customers for idea evaluation. In cases where the target population is rather small, this might constitute a challenge. Examining efficient procedures of how to select the best ideas including who should ideally be involved (experts vs. consumers) could be a fruitful line for future research.

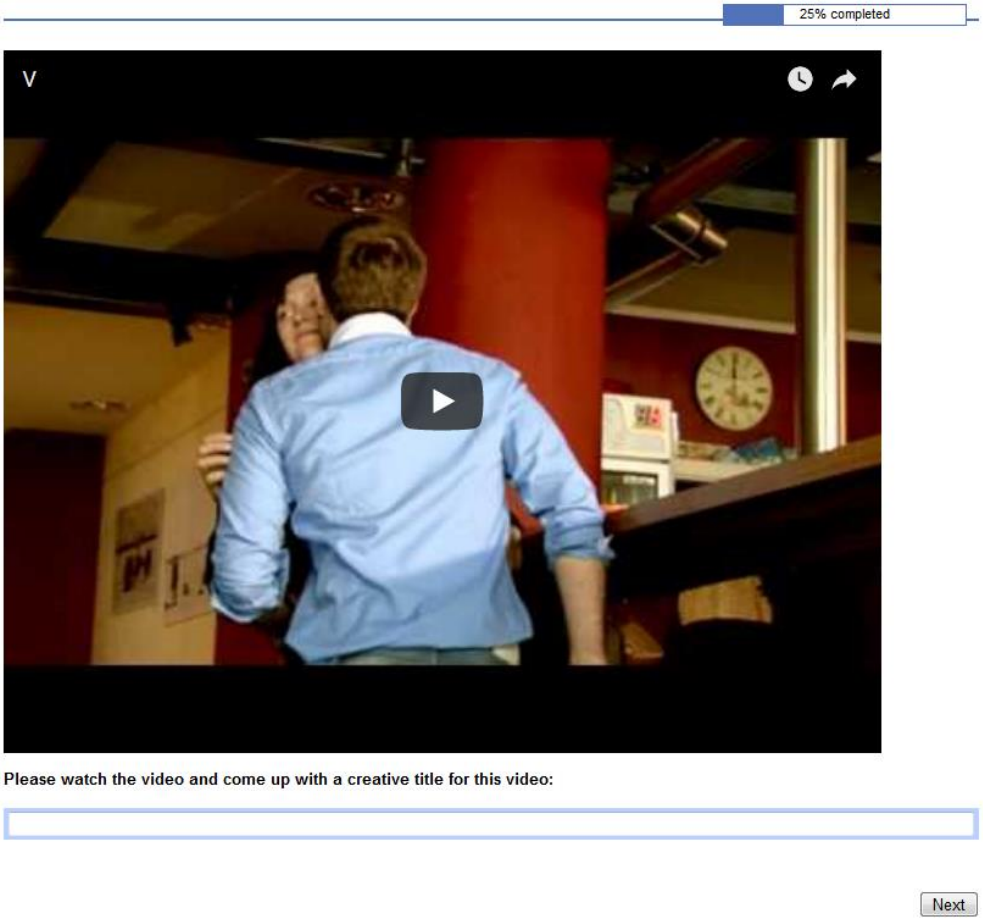
3.5 Appendix to Chapter 3

A Instructions (for Experts)



In the following you will see a short video (72 seconds). We ask you to come up with a title, which we can use to upload this video on our homepage. The aim is to find a creative title that generates high click rates and is related to the video. We will collect several titles from different people. The best title is the title that generates the highest click rates among people from the internet marketplace MTurk.

You will receive $60 for performing this task. Additionally, the person who comes up with the best title will receive a reward of $400.



Please watch the video and come up with a creative title for this video:

B List with all titles and their scores

TABLE 3.4: ALL TITLES WITH SCORES

Rank	Quality	Expert	Title
1	0.61	1	Liars, Cheats & Theives: The Truth about Human Character
2	0.60	0	Immoral Statistics: Breaking Society's Rules
3	0.55	0	Cheating, the new social norm?
4	0.52	0	Frappuccinos, fraud, and fucking.
5	0.45	0	See How Society is Stealing Your Hard-earned Money
5	0.45	0	HAND IS QUICKER THEN THE EYE
7	0.43	1	People Suck -- Facts about Liars, Cheats, and Fraud
7	0.43	0	You won't believe these shocking facts about cheaters!
9	0.40	1	REVEALED: The Uncomfortable Truth About People you TRUST
9	0.40	0	Cheating scandals that will SHOCK you!
9	0.40	1	Fraud, Lies, Deception! - The Hidden Costs of Cheating
9	0.40	0	Living in a Society of Liars: Do You Really Know the Truth?
13	0.38	0	Flirting with Danger
14	0.38	0	Cheaters Among Us and The Stunning Statistics-Don't Get Fooled Again
14	0.38	0	Four facts about cheating and fraud, number two will surprise you
14	0.38	0	Hidden truths in american society.
14	0.38	1	Character - what you do when no one is looking. Which side of the fence are you on?
18	0.35	1	Con-Science: The Math of Bad Behavior
18	0.35	1	Cheating: Human Nature or Cultural Phenomenon?
18	0.35	1	Cheaters Think Everyone Cheats
18	0.35	0	A Cheat Sheet
22	0.33	1	Is cheating a natural human instinct?
22	0.33	0	Cheaters everywhere among us
22	0.33	0	Sick Sad World
22	0.33	0	Is cheating the norm?
22	0.33	0	Mass Deception and Fraud
22	0.33	1	Humans lie. Use protection.
22	0.33	0	Cheats Rule The World
22	0.33	1	Caught!! These Statistics on Cheating Will Blow Your Mind...
30	0.32	1	First You See It, Then You Don't!
30	0.32	1	Don't Get Swindled Or Cheated On. Catch Them In The Act.
32	0.30	1	Do cheaters ever prosper?
32	0.30	0	Lying, Cheating, & Stealing: The Fabric of Humanity.
32	0.30	0	The lie in my pocket
32	0.30	0	Liars, Cheats, and Deceits
32	0.30	0	Can You See Yourself in This Video?
32	0.30	1	Think People Are Honest? Think again...
32	0.30	0	Everyone's a Fraud
39	0.29	0	Cheating Analytics
39	0.29	0	Fake It 'Til You Make It - Everyone Else Does! The Truth About Liers, Cheaters and Thieves.
39	0.29	0	The Fraud of Human
39	0.29	0	Society of Cheaters
39	0.29	0	Cheaters Gonna Cheat: The New Epidemic
44	0.28	0	Circle of Lies: To Cheat or Not to Cheat?
44	0.28	0	What do students, baristas and cheating husbands have in common? The answer may surprise you.
44	0.28	1	4 Proven Reasons You Should Trust No One

47	0.26	0	Facts about CHEATING!
47	0.26	0	Can you spot the cheats and the lies all around?
47	0.26	0	Sugar Plums or Cheating Liars: Who Can You Trust?
47	0.26	0	Cheating - Believe it or Not..A lot of us do it.
51	0.25	0	Den of Thieves: The Cheaters, Frauds and Liars You Should Beware
51	0.25	0	Cheaters Bar & Lounge: We always look the other way!
51	0.25	0	Is He Cheating?
51	0.25	0	Who's afraid of fraud?
51	0.25	1	IT'S HAPPENING RIGHT UNDER YOUR NOSE
51	0.25	0	Dirty Cheaters
51	0.25	0	Cheating makes the world go round
58	0.24	0	Some who seem sweet may also cheat
58	0.24	0	The world of cheating
58	0.24	0	A mountain of deception
58	0.24	1	The Data-Backed Truth About Cheating In Our Society
58	0.24	0	Surrounded by Deception
58	0.24	0	A Fraud-Ean world.
58	0.24	0	Cheaters beware! You will be caught in the act
58	0.24	0	Cheating: A one way ticket to ruining your life, losing your wife, and living in strife
58	0.24	0	Morals Down the Drain: a Digital Ballet
67	0.22	0	Go ahead, Cheat you'll get caught.
67	0.22	0	Dishonest Drinkers
69	0.21	0	Cheating: A Part of Human Nature?
69	0.21	0	Cheaters Never Prosper (Unless They Get Away With It)
69	0.21	0	Just another day cheating and stealing.
69	0.21	0	Do cheaters and thieves really win?
73	0.20	0	What are the odds the people in your life are cheating? Come find out!
73	0.20	0	Cheating: a highly indulgent activity
73	0.20	0	Cheating Facts Set to Upbeat Christmas Music
73	0.20	0	Creepy Cheaters.
73	0.20	0	Find Out 5 Things About Cheating You Didn't Already Know and Have Your Mind Blown.
73	0.20	0	Do you have a favorite thief in your life?
73	0.20	0	Everybody cheats... That makes it okay, right?
73	0.20	1	Is Someone Lying to You Right Now? Probably.
73	0.20	1	Cheats & Deceits
82	0.19	0	Lie, Cheat and Steal
82	0.19	0	People in your life are probably cheaters.
82	0.19	1	Dishonest Workers: You won't believe how much they're costing businesses
82	0.19	0	Cheating Statistics
82	0.19	0	Cheating- is it worth the cost?
82	0.19	0	Cheating: Who Is Doing It?
82	0.19	0	greatest crimes
82	0.19	0	Fraud is ruining corporate profits
90	0.16	0	Decepticons among us
90	0.16	0	The wide world of cheats and frauds
90	0.16	1	Life's a cheat:The illusion of honesty
90	0.16	1	A Society of Cheaters
90	0.16	1	$2.9 Trillion Made By CHEATING Every Single Year!
90	0.16	1	that Add Up Big
90	0.16	0	Ethically Unbound: The Numbers Behind Bad Behaviors

97	0.15	0	Quick Facts about Integrity
97	0.15	0	Beg, Borrow or Steal: The Big Lie & What You Should Know About Everyone Close To You.
97	0.15	1	Fraud: Who pays the cost?
97	0.15	0	Deception All Around Us: Who Can You Trust?
97	0.15	0	Cheating and you: Be the person, not a worseon.
102	0.14	0	Deceptively educational: Think you know all there is to know about cheating?
102	0.14	0	Watch This If You Wanna Know How Many
102	0.14	0	Fraud and Cheating, what is the cost?
102	0.14	0	Cheater Cheater, Deceitful Meter
102	0.14	0	cheating and its relevance
102	0.14	0	Learn Cheating and Fraud Statisitcs While Listening to a Techno Version of Dance of the Sugar Plum Fairies!
108	0.11	0	You too may be contributing to a world of uncertainty
108	0.11	0	Waiting for luck? Just Cheat. If you got away, it's because of hard work.
110	0.10	0	Is cheating the new normal?
110	0.10	1	A Cheat Sheet to Cheat
110	0.10	0	Cheating, Fraud and your spouse. Why it's affecting YOU
110	0.10	1	If Someone's Calling Your Bluff, the Joke May Be on You
110	0.10	1	Thieves, Liars & Cheats All Around Us
110	0.10	1	Visit a pub and witness the darker side of human nature!
110	0.10	0	So Many Suckers
117	0.06	0	Find out if you are one of these cheaters
118	0.05	0	ALL KINDS OF CHEATERS
118	0.05	0	We know you cheat!
118	0.05	0	Video infographic on cheating

C Descriptive Statistics

TABLE 3.5: SAMPLE CHARACTERISTICS

	Non-experts				**Experts**				**Difference**
	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	*p*-value
Female	0.36	0.48	0	1	0.57	0.50	0	1	0.047
Age	33.06	8.73	26.34	70	40.57	11.39	26	67	0.001
White	0.69	0.47	0	1	0.90	0.31	0	1	0.021
Education[a]	1.53	0.71	0	3	2.27	0.64	0	3	0.000
Expertise (self-rated)	3.11	1.14	1	5	4.63	0.56	3	5	0.000

Notes: The sample size is 89 for non-experts as one worker did not respond to the questions and 30 for experts. The p-values are based on two-sided Mann-Whitney test statistics. [a] We transferred questionnaire data on education into qualitative ascending categories that are in accordance with years of education no school diploma=0, 1=high school diploma, 2=bachelor's degree; 3=master's degree or higher.

Chapter 4

Do Fines Deter Unethical Behavior? The Effect of Systematically Varying the Size and Probability of Punishment

4.1 Introduction

Unethical behavior is prevalent in many economic transactions, and understanding how to deter it is important. Even when the unethical behavior occurs on a small scale, the aggregate effect on the economy could be large. Consider, for example, illegal downloads of music. According to the RIAA report (www.riaa.com/faq.php), from 2004 through 2009, approximately 30 billion songs were illegally downloaded from file-sharing networks. An Institute for Policy Innovation (www.ipi.org) report concludes global music piracy causes $12.5 billion in economic losses every year, 71,060 US jobs lost, and a loss of $2.7 billion in workers' earnings. Or consider the case of employee theft in the workplace, which has been estimated to result in up to $200 billion in losses per year for US companies (Murphy 1993), with $15.9 billion in the retail industry alone (2008 National Retail Security Survey). Although in both of these examples each offender may only steal a small amount of money, due to the number of people engaging in such behavior, the collective damage to the industry is huge.

The standard economic model of deterring criminal behavior (Becker 1968) assumes the decision of whether to commit a crime is based on expected utility. Three outcomes are possible: the decision maker can choose not to engage in crime. If she chooses to commit a crime, either she is successful or she is audited and punished. In this model, the decision maker weighs the utility of each outcome and the associated probabilities, and commits the crime if the expected utility of doing so is higher than that of not committing the crime. A policy aimed at deterring decision makers from committing a crime could then be based on increasing the probability of catching the criminal or on changing the gains and losses associated with the outcomes. In this paper, we experimentally test how such policies affect unethical behavior.

The question of how the size and likelihood of punishment affect deterrence originated in the law and criminology literature. This work dates back to Beccaria (1764), who advanced the idea that "crimes are more effectively prevented by the certainty than the severity of punishment." Since then, a large empirical work in different fields has tried to estimate the effects of the certainty and severity of punishment on "real" crimes, for example, by studying the effect of changes in police force or changes in penalties. Overall, this work predominantly finds that increases in police manpower are correlated with reductions in crime, whereas the effect of the harshness of sanctions is smaller.[1] Chalfin and McCrary (2017) discuss such results in their review of this literature, and point out the challenges in identifying causal effects, for example, because changes in police force or sanctions are typically not random. Another challenge in this literature is the proxies used to identify the probability of being detected, which is inferred from past data.

In this paper, we experimentally study choices of a smaller scale than the empirical literature on crime. In particular, we are interested in how the probability of punishment and its size affect ethical decision-making. Most of us do not commit large-scale crimes, but face ethical decisions, such as whether to download music illegally or to lie to achieve an advantage. On a smaller scale, these decisions are very common. For example, DePaulo and Kashy (1998) estimate that, on average, people lie twice a day.

To study deterrence of small-scale unethical behavior, we have our experimental participants play a deception game (Gneezy 2005) in which they receive private information regarding payments, and are asked to send a message to another player. In the game, participants know that sending a false message increases the chance of a higher payoff to the Sender at the expense of the other participant. To identify what kind of policy would most effectively deter deception in this game, we use a systematic approach based on Becker's model. In 20 different treatments, we

[1] See Levitt (1997, 2002), Evans and Owens (2007), Lin (2009), and Chalfin and McCrary (2013) for results on the effect of police manpower on crime. For sanctions, the results are mixed. The early literature investigating the effects of sentence enhancement on specific crimes, such as gun crimes, found few deterrence effects (see, e.g., Loftin and McDowall 1981, McDowall, Loftin, and Wiersma 1992). However, more recent work focusing on the effect of changes in the sanction regime found some impact of harsher penalties (see, e.g., Drago, Galbiati, and Vertova 2009, Helland and Tabarrok 2007), though these effects are small in magnitude.

vary the probability of and the fine associated with being audited. We also investigate the effect of making the probability and the fines ambiguous in this one-shot game.

As explained above, the standard model of deterrence predicts individuals to be sensitive to decreases in the expected payoffs that arise both from an increased likelihood of being audited and from higher penalties if detected. Under expected utility, for small payoffs such as the ones used in our experiment, the utility function is almost linear, and the expected value is a good approximation of behavior (Rabin 2000). These predictions can be generalized to other models of decision under uncertainty, such as prospect theory (Kahneman and Tversky 1979). However, in prospect theory, small probabilities, such as 0.05, are overweighted relative to medium probabilities, such as 0.50. Hence, under prospect theory, participants are still predicted to be sensitive to changes in probabilities, but less sensitive relative to expected utility.

In the one-shot experiments, when presented with only one set of parameters, we find that participants are sensitive to the size of the fine, but are insensitive to changes in the probability of being fined. This result suggests that, in our data, the decision to lie is not based on calculations of the expected utility of a given sanction, potentially due to the complexity of making such calculations. Instead, individuals seem to base their choices to lie on simple decision heuristics that depend on the magnitude of fines, which may be more salient and easier to understand than detection probabilities.[2]

This result is in contrast to experimental findings in the tax evasion literature, which, starting with Friedland, Maital, and Rutenberg (1978), investigates how the probability and size of punishment affect the decision to evade taxes in laboratory simulations testing the deterrence model proposed by Allingham and Sandmo (1972). Much of the subsequent work finds tax-evasion decisions are sensitive to changes in expected costs determined by increases in detection probabilities (Friedland 1982, Spicer and Thomas 1982, Webley 1987, Alm, McClelland, and Schulze 1992, Beck, Davis, and Jung 1991, Bott 2016). Similar results have been found outside of the domain of tax evasion, in experiments designed to test deterrence in allocation decisions (Schildberg-Hörisch and Strassmair 2012, Khadjavi 2015, Harbaugh, Mocan, and Visser 2013), public goods games (Anderson and Stafford 2003), or studying other types of incentivized choices in the lab (DeAngelo and Charness 2012).

[2] This result is related to the broader literature on bounded rationality, which shows that, in complex environments, decisions are often not based on calculations of expected utility, for example due to cognitive limitations or limited attention, but rather rely on heuristics or cognitive shortcuts (see Simon 1955). For empirical and theoretical work on these biases see e.g., Tversky and Kahneman, 1973, Chetty, Looney and Kroft 2009, Bordalo, Gennaioli and Shleifer 2012, 2013, Schwartzstein 2014, Enke and Zimmermann forthcoming, Gabaix 2017.

These experiments, however, used either a within-participant design in which individuals evaluated the magnitude of detection probabilities by comparing different detection likelihoods, and/or a repeated-game design in which participants made repeated decisions with feedback.[3] In such designs, incorporating detection probabilities into the decision process may be easier. Yet, these settings are profoundly different from settings in which individuals only face one given detection probability where feedback is rare as captured by our one-shot between-participant design.

To better understand whether complexity associated with evaluating probabilities in isolation can explain why participants were not responsive to changes in probabilities in our experiment and understand the discrepancy with previous results, we conduct two additional experiments. First, we have participants play a variation of the one-shot experiment in which they indicate their decision to lie for all four possible detection likelihoods. We find that, on top of being sensitive to fines, in a setting where individuals can directly compare different probability levels to each other, participants become sensitive to changes in detection probabilities. We conjecture this sensitivity arises from the fact that presenting individuals with several probabilities reduces complexity by providing them with a reference point against which they can compare and contrast their decisions. As a result, the magnitude of different probability levels becomes more salient.

Second, we examine data from an additional experiment built on a growing literature contrasting risky choices under *descriptions* versus *experience* of probabilities (see, e.g., Barron and Erev 2003, Hertwig, Barron, Weber, and Erev 2004, Hertwig and Erev 2009). We examine the dynamics of lying behavior in a repeated setting where individuals can experience being monitored after every round and receive immediate feedback regarding punishment. In each round, they face the same detection probability. In such a setting, even if decision makers have limited computational abilities and do not rely on calculations of expected utility, they can directly experience what a given detection probability means in terms of audit frequencies. As in the one-shot experiment, we find that in the first round, individuals are completely insensitive to probabilities but are sensitive to the size of the penalty. However, over time, lying behavior becomes responsive to increases in the detection probability as well.

Taken together, the results of our two additional experiments reconcile our findings with the previous literature. Even though participants do not use all the information available but rather use

[3] In Table 4.6 in Appendix A we summarize this work and classify it according to the experimental methods adopted (repeated vs. one-shot experiment; between- vs. within-participant design).

cognitive shortcuts based on the size of the fine when evaluating probabilities is complex, they become responsive to probabilities when they can directly compare probabilities to each other or directly experience audit frequencies over time. Our findings suggest that policy interventions that introduce harsher fines are likely to be a successful means of deterrence. Further, policies designed to target detection probabilities may be effective when feedback is frequent, such as in the domain of fare evasion in public transportation. However, for unethical behavior in which individuals receive only rare feedback, such as small-scale tax evasion, presenting individuals with descriptions of detection probabilities is likely to be ineffective unless changes in the magnitude of such probabilities are made easy to evaluate.

The remainder of the paper is organized as it follows. Section 4.2 presents the experimental design, section 4.3 the predictions, and section 4.4 discusses the results of the one-shot data. In section 4. 5, we present the design and results of the experiment in which lying occurs over several rounds. Section 4.6 concludes.

4.2 Experimental Design

4.2.1 The Game

Unlike previous work, where the decision to "misbehave" does not capture ethical considerations, we design our experiment to capture unethical behavior in a setting where maximizing profit requires individuals to lie, rather than to make choices over monetary allocations, and involves harming another participant. We use a version of the deception game (Gneezy 2005, Dreber and Johannesson 2008, Sutter 2009, Erat and Gneezy 2012, van de Ven and Villeval 2015) in which participants can lie to increase their earnings at the expense of another participant. The experiment involves two players, a Sender and a Receiver. The Sender starts the experiment with a $10 participation fee and is privately informed about the outcome of a 10-sided die roll. The Sender is then asked to send a message regarding the die roll result to the Receiver. She can choose one of 10 possible messages that state, "The outcome of the die roll is x", where x is a number between 1 and 10.

After observing the message, the Receiver is asked to choose a number between 1 and 10. The Sender's message is the only information the Receiver has about the outcome of the die roll; that is, the Receiver does not know the actual number that came up in the die roll. The Receiver's choice determines the payoffs of both players. Importantly, the Sender is informed about the payoffs associated with the Receiver's choice, whereas the Receiver is not. The Receiver only knows the Sender has private information about the actual die roll result and that the payment for both players depends on whether her choice corresponded to the actual outcome of the die roll.

Our implementation of the game has two different payment options. If the Receiver chooses the number corresponding to the actual outcome of the die roll, the Sender earns $5 (on top of the participation fee) and the Receiver earns $15. If the Receiver chooses a number other than the actual die roll, the Sender earns $15 (on top of the participation fee) and the Receiver earns $5. These payoffs are in addition to a $0.25 base payment for taking part in the experiment. Therefore, in this game, the Sender has an incentive to lie by sending a false message to the Receiver. Note the Receiver does not know hers and the Sender's incentives - not even that they are not aligned.

4.2.2 Treatments

The baseline treatment does not include monitoring. In the remaining treatments, the Sender is informed that some messages will be audited. If audited and the Sender's message does not correspond to the actual outcome of the die roll (i.e., if the Sender lies), the Sender is fined, losing a fraction of the total earnings. We systematically vary the size of the fine and the probability of being audited. We cross two known fine levels with four different probabilities of being audited, both between participants (section 4.2.2.1) and within participants (section 4.2.2.2). In Appendix C, we also extend our investigation to unknown fines levels and/or probabilities to study how ambiguity affects responses to auditing in a between-participant design.

4.2.2.1 Varying Fines and Probabilities: Between-Participant Treatments

In the first eight treatments, we use a between-participant design in which we vary the size of the fine to be either high or low, and cross it with four different probabilities of being audited.

TABLE 4.1: EXPERIMENTAL TREATMENTS

Treatment	Size of the fine	Probability of being audited	N
PANEL A	(between participants)	(between participants)	
NO AUDITING			
Baseline	-	-	151
AUDITING			
LowF_0.05	Low	0.05	149
LowF_0.10	Low	0.10	151
LowF_0.25	Low	0.25	156
LowF_0.50	Low	0.50	149
HighF_0.05	High	0.05	154
HighF_0.10	High	0.10	149
HighF_0.25	High	0.25	149
HighF_0.50	High	0.50	152
PANEL B	(between participants)	(within participants)	
LowF_pWithin	Low	0.05 0.10 0.25 0.50	95
HighF_pWithin	High		96
PANEL C	(within participants)	(between participants)	
Fwithin_0.05	Low High	0.05	108
Fwithin_0.10	Low High	0.10	104
Fwithin_0.25	Low High	0.25	105
Fwithin_0.05	Low High	0.05	108
		Total	**1976**

Size of the Fine—We vary the size of the fine to either $12.50 (Low Fine) or $25 (High Fine). In the Low Fine treatments, the Sender is informed that if audited and the message she sent to the Receiver does not correspond to the actual outcome of the die roll, she will lose half her earnings. In particular, she will lose half the $10 participation fee and half the payment associated with the die roll experiment, ending up with a payment of $12.50 for the experiment (on top of the $0.25 base fee). In the High Fine treatments, the Sender is informed that if audited and the message does not correspond to the actual die roll result, she will lose all her earnings and only receive the $0.25 base fee.

Probability of Being Audited—We vary the probability of being audited p to be equal to 0.05, 0.10, 0.25, or 0.50 between participants. The eight resulting treatments are displayed in Panel A of Table 4.1.

An important policy question concerns the relative effectiveness of changes in expected costs that come from changes in penalties as opposed to changes in detection probabilities. To investigate this question, we need to exogenously vary both fines and probabilities while keeping constant the expected costs of deterrence. Only a few studies vary fines and probabilities in the realm of a single experiment. Previous experimental work has found mixed results, with some findings suggesting a greater deterrent effect of sanctions than detection probabilities (Friedland, Maital, and Rutenberg 1978, Anderson and Stafford 2003, Friesen 2012), other research suggesting the opposite pattern (Friedland 1982, Webley 1987), and additional work suggesting that different combinations of fines and probabilities are equally effective and can therefore be considered substitutes (Schildberg-Hörisch and Strassmair 2012, Khadjavi 2015)[4].

Our experimental setup provides us with an opportunity to cleanly investigate the relative effectiveness of fines and auditing probabilities in a between-participant design, by comparing the behavior of participants that face identical expected payoffs from lying but are exposed to different combinations of fines and probabilities (treatments HighF_0.05 and LowF_0.10; and HighF_0.25 and LowF_0.50 in Table 4.1).

4.2.2.1 Varying Fines and Probabilities: Within-Participant Treatments

Changing the Probability of Being Audited within Participants—In the next two treatments, we fix the size of the fine and vary the probability of being audited within participants. The Sender indicates the message she wants to send to the Receiver for each possible probability level of being audited (p=0.05, p=0.10, p=0.25, and p=0.50). In the LowF_pWithin, the fine is $12.50, and in HighF_pWithin it is $25. Participants are informed that one of their choices will be selected at random and used for payment. Panel B of Table 4.1 describes the LowF_pWithin and HighF_pWithin treatments.

Changing the Size of the Fine within Participants—In four additional treatments, we fix the probability of being audited and vary the size of the fine within participants to either $12.50 or $25. In these treatments, the Sender is asked to indicate the message she wants to send for both possible fine levels. The probability of being audited is 0.05 in Fwithin_0.05, 0.10 in Fwithin_0.10, 0.25 in Fwithin_025, and 0.50 in Fwithin_0.50. Again, participants are informed

[4] Most of this work adopts within-participant designs, which may lead to misleading interpretation of the relative effectiveness of fines and detection probabilities.

that one of their choices will be selected at random and used for payment. Panel C in Table 4.1 illustrates these treatments.

4.2.3 Procedure

We conducted the experiment on Amazon's Mechanical Turk (mTurk), one of the largest online marketplaces for task-based work. mTurk is a growing online platform that has been widely used to conduct experiments in the social sciences (see Horton, Rand, and Zeckhauser 2011, and Paolacci, Chandler, and Ipeirotis 2010) and it has been recently used in economics, for example, to study redistribution preferences (Kuziemko, Norton, Saez, and Stantcheva 2015) or the effect of monetary and non-monetary incentives on effort provision (DellaVigna and Pope 2018). We recruited 2,731 workers who were randomly assigned to the different treatments reported in the main text and the Appendix. The task was posted on mTurk as a study on decision-making, with a description stating that participants would earn $0.25 plus bonus opportunities that would last about 5-10 minutes. Participants could take up to one hour to finish the experiment. All instructions are posted in Appendix F. We took precautionary measures to prevent participants from taking part in this experiment multiple times. Specifically, before the participants could see the instructions, they had to type in their worker ID.[5] Additionally, we took several standard mTurk measures to ensure quality. First, we restricted the sample of workers to those with US residency; that is, the task was only shown to workers with a US address. Second, to exclude automatic robots, only workers with a past approval rate of 95% could take part in the study, and we asked participants to enter an individual completion code at the end of the task to be eligible for payment.

After reading the consent form, participants learned that on top of their $0.25 fixed payment, 1 out of 20 participants would be selected randomly and paid a monetary bonus according to the instructions. We informed participants in the role of Sender that, if selected, they would receive a $10 participation fee and would be matched randomly with another mTurk worker, the Receiver, who would receive a $0.25 fixed payment. Then, participants in the role of Sender were randomly assigned to one of the 20 treatments. The instructions were presented on one screen that included the description of the task and the payoff for both the Sender and the Receiver. On this page, they were asked to choose the message for the Receiver. In the baseline treatment, the instructions did not mention auditing. We informed the Sender that the number the Receiver chose would

[5] Six workers, however, circumvented these precautionary measures and were able to participate in the experiment twice. We only look at the first participation. Excluding these workers altogether does not change the results.

determine her payment in the experiment. In the remaining treatments, participants learned about the chance of being audited. We informed them we would randomly select one participant out of x, where x was 20, 10, 4, or 2, depending on the treatment, and check their message. If the message did not correspond to the actual outcome of the die roll, participants would be fined either half or all their earnings from sending a false message, depending on the treatment. We also had ambiguity treatments in which fines and probabilities were ambiguous (see Table 4.7 in Appendix C). In these treatments, we informed them that some participants would be randomly audited and/or that if fined, they would lose some of their earnings.

Once the Sender clicked on the message she chose to send, she was directed to another screen where we asked her to fill out a survey and report some demographics such as gender, age, nationality, and field of study. She was then provided with a unique code to be entered on mTurk for payment purposes.

After collecting all the data for the Sender, we separately recruited 136 mTurk workers (46% women) who participated in the experiment in the role of Receiver in exchange for a $0.25 fixed payment and the opportunity to receive additional earnings. After reading the instructions, the Receiver learned about the message from the Sender and had to make a choice. 65% of Receivers followed the message, which suggests sending a false message in this context had a high chance of penalizing the Receiver.[6]

4.3 Predictions

We follow Becker (1957) and the standard assumptions he made (in particular, no moral costs of lying). The agent's decision depends on the costs and benefits associated with deception. The expected utility from lying is

$$EU_j = p_jU_j\,(Y_j - f_j) + (1- p_j)U_j(Y_j)$$

where Y_j is the payoff from lying and f_j is the fine if audited and caught lying. In the model, an increase in p_j or f_j will reduce the expected utility of deception. Depending on the agents' risk preferences, an increase in p_j would either result in higher, identical, or lower expected utility from lying, as compared to an equivalent percentage increase in f_j. According to Rabin (2000), for small stakes, such as the ones in our experiment, the utility function is close to linear. Therefore, under expected utility, deceptive behavior is predicted to approximately reflect changes in expected

[6] The men followed the message in 60% of the cases; the women did in 73%. This difference is not significant (χ^2=2.26, p<.133).

payoff as compared to the sure payoff when telling the truth.

For the predictions of Prospect Theory, we assume the reference point to correspond to the decision maker's participation fee plus his/her earnings from telling the truth. This way, if the decision maker lies and is (not) caught, he/she would end up in the domain of losses (gains). Under these assumptions and the standard parameters in the literature, for a high fine, decision makers would be sensitive to changes in probabilities that range between 0.05 and 0.5, and would switch from lying to telling the truth when p is equal to 0.25 (*HighF_0.25*). For a low fine, individuals would always choose to lie regardless of the detection probabilities. See Appendix B for the calculations.

TABLE 4.2: EXPECTED PAYOFFS FROM LYING

	Probability of being audited			
	0.05	0.1	0.25	0.50
Low Fine	$24.38	$23.75	$21.88	$18.75
High Fine	$23.75	$22.50	$18.75	$12.50

Table 4.2 reports the expected payoffs of lying for all possible combinations of fines and probabilities. The Sender's expected payoff from lying decreases with increases in both the probability of being audited and the size of the fine. In two of the cases we study, the expected payoffs from lying are identical but determined by a different combination of fines and probabilities (HighF_0.05 and LowF_0.10; HighF_0.25 and LowF_0.50). These two combinations allow us to test whether our participants are more sensitive to the probability of being audited or to the size of the fine. The null hypothesis under the predictions discussed above is that we would find no difference in lying behavior when the expected payoff is the same and agents are risk neutral. Under Cumulative Prospect Theory (CPT), individuals tend to over-weight small probabilities. They would therefore be less sensitive to changes in expected value that come from variations in detection probabilities than to changes that come from variations in the size of the fine. For the parameters of our experiment, CPT predicts equal lying rates in HighF_0.05 and Low_0.10, but not in HighF_0.25 and LowF_0.50.

The above predictions are based on the assumption that participants would be indifferent between outcomes that requires them to lie and identical outcomes that are reached by telling the truth. However, recent work finds people have lying costs (e.g., Gneezy 2005, Charness and Dufwenberg 2006, Belot, Bhaskar, and van de Ven 2012, Battigalli, Charness, and Dufwenberg

2013, Cappelen, Sørensen, and Tungodden 2013). Some choose to tell the truth and lose potential earnings even in the absence of monitoring. Based on this fact, lying rates throughout our treatments may be smaller than predicted by expected utility or CPT under the assumption of no lying costs.

4.4 Results

A total of 2,731 (53.1% female) participants completed the experiment, including those who participated in the treatments reported in Appendix C (*N*=755). The average age was 36 (SD=11.69, Median=33). In term of ethnicity, 76.6% of the participants indicated being Caucasian, 7.03% indicated being African-American, 6.31% reported being Asian, and 4.51% reported being Hispanic. The remaining participants either indicated a mixed ethnicity or did not report their ethnicity. The wide majority of participants (97.47%) indicated English was their first language. We asked the remaining participants to report their English proficiency level on a scale from 1 to 7; the average score was 6.71 (SD=.72). In this section, we first present the results of the between-participant treatments and then move to the within-participant treatments. In Appendix C, we also explore how ambiguity in the deterrence parameters affects lying behavior.

4.4.1 Varying Fines and Probabilities: Between-Participant Data

Table 4.3 displays the fraction of lies by probability of being audited across treatments. In the baseline with no auditing, 51.7% of the participants (*N*=151) sent a false message to the Receiver.

Result 1 *Introducing auditing decreases the frequency of lies.*

Our first result is evident from the Figure 4.1: when we introduced auditing, participants reacted by lying significantly less often ($p \leq .002$ for all treatments, chi-squared test). This result is in line with work showing that introducing monitoring and punishment works in deterring behaviors that negatively affect others (e.g., Khadjavi 2018).

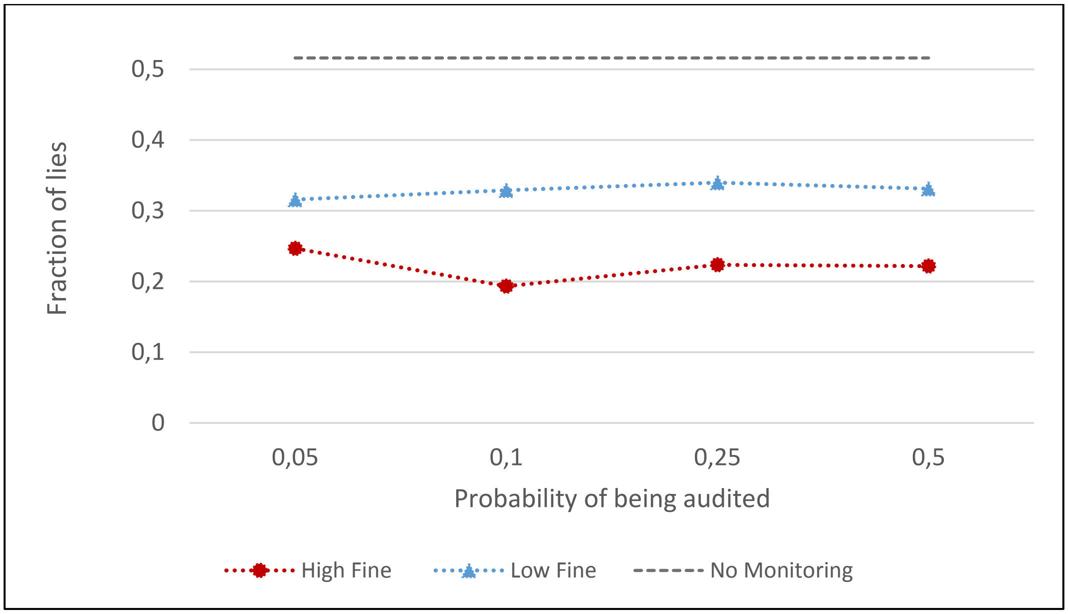


FIGURE 4.1: FRACTION OF LIES BY PROBABILITY OF BEING AUDITED AND BY FINE SIZE (BETWEEN-PARTICIPANT TREATMENTS)

As predicted, we find participants were sensitive to fines. For any given probability, when the fine was high, participants lied significantly less than when the fine was low (*p*<.001, chi-squared test). Hence, we conclude that in our data participants are sensitive to increases in the size of a penalty associated with deceiving.

Result 2 *Deception rates decrease when the size of the fine increases*.

We further find participants did not react to substantial increases in the probability of being audited. As Table 4.3 shows, when the fine for being audited was low, the fraction of participants who lied when the probability of being audited was 5% was .315. This fraction was .342 when the probability increased to 50%; this difference is not statistically significant (χ^2=.243 *p*=.622, chi-squared tests, *p*≤.622 for all other pairwise comparisons). We find a similar pattern for the high fine: a fraction of .24 participants lied when the probability was 5%, and this fraction was not significantly lower when the probability was 50% (.23, (χ^2=.114, *p*=.735), all pairwise comparisons for the high-fine treatments are not significant (*p*≤.274, chi-squared test).

Result 3 *Participants are not sensitive to the probability of being audited.*

TABLE 4.3: DESCRIPTIVE STATISTICS OF LYING RATES

PANEL A	Probability between participants – Fine between participants						
NO AUDITING							
Treatment	**Mean**	**N**					
Baseline	.5166	151					
AUDITING	**Low Fine**			**High Fine**			
Probability	**Treatment**	**Mean**	**N**	**Treatment**	**Mean**	**N**	**p-value** (χ^2-test)
0.05	LowF_0.05	.315	149	HighF_0.05	.247	154	0.184
0.10	LowF_0.10	.331	151	HighF_0.10	.195	149	0.007
0.25	LowF_0.25	.340	156	HighF_0.25	.228	149	0.031
0.50	LowF_0.50	.342	149	HighF_0.50	.230	152	0.032
Overall	All Low Fine treatments	.332	605	All High Fine treatments	.225	604	0.000
PANEL B	Probability within participants – Fine between participants						
0.05	LowF_pWithin	.442	95	HighF_pWithin	.198	96	0.000
0.10	LowF_pWithin	.379	95	HighF_pWithin	.146	96	0.000
0.25	LowF_pWithin	.274	95	HighF_pWithin	.094	96	0.000
0.50	LowF_pWithin	.274	95	HighF_pWithin	.031	96	0.000
Overall	LowF_pWithin	.318	95	HighF_pWithin	.117	96	0.000
PANEL C	Probability between participants – Fine within participants						
0.05	Fwithin_0.05	.324	108	Fwithin_0.05	.194	108	0.002
0.10	Fwithin_0.10	.317	104	Fwithin_0.10	.163	104	0.000
0.25	Fwithin_025	.343	105	Fwithin_025	.171	105	0.000
0.50	Fwithin_0.50	.426	108	Fwithin_0.50	.194	108	0.000
Overall	Fwithin_0.05 Fwithin_0.50	.353	425	Fwithin_0.05 Fwithin_0.50	.181	425	0.000

Table 4.4 reports this result using an OLS regression analysis exploring the effects of the size of the fine and of different probabilities of being audited on lying. Column 1 shows an increase in the size of the fine decreased the probability of lying by 9.9 percentage points (*p*=.020). Increasing the probability of being audited had no effect on lying. We also find no significant interaction between fine and probability, suggesting the effect of fines on lying did not depend on the probability of being audited. In the control condition with no auditing, participants lied significantly more often than in all other treatments in which auditing was introduced (*p*<.001 for all comparisons). Column 2 shows that controlling for gender does not change the results. We also find women lied significantly less often than men (*p*<.001). This result is in line with previous findings on gender differences in the propensity to tell selfish black lies (Dreber and Johannesson

2008, Erat and Gneezy 2012). We observe no evidence that women react differently than men to fines and probabilities (see Table 4.9 in Appendix D).

Further, we find these results are robust to controls such as age (column 3) and ethnicity (column 4). Excluding participants who said they had difficulty understanding the instructions (*N*=152, 11.2%) or excluding those who did not answer the post-experimental comprehension check correctly (*N*=326, 23.8%) strengthens the results (see Table 4.10 and 4.11 in Appendix D). Finally, in Table 4.12 in Appendix D we report the results of an OLS regression in which we include dummy variables for each of the between-participant treatments, and find similar results.

TABLE 4.4: EFFECT OF FINE AND PROBABILITY ON LYING

Dependent variable:	Lied (1=Yes)			
	(1)	(2)	(3)	(4)
High Fine	-0.099**	-0.098**	-0.101**	-0.100**
	(0.042)	(0.041)	(0.041)	(0.041)
Probability	0.049	0.053	0.065	0.064
	(0.110)	(0.109)	(0.108)	(0.108)
High Fine*Probability	-0.038	-0.054	-0.061	-0.072
	(0.147)	(0.146)	(0.146)	(0.147)
No monitoring	0.195***	0.199***	0.199***	0.201***
	(0.051)	(0.051)	(0.050)	(0.050)
Female		-0.106***	-0.089***	-0.091***
		(0.025)	(0.025)	(0.025)
Age			-0.004***	-0.004***
			(0.001)	(0.001)
Ethnicity	No	No	No	Yes
Constant	0.321***	0.377***	0.515***	0.503***
	(0.031)	(0.034)	(0.049)	(0.051)
Treatments	LowF_0.05- HighF_0.50	LowF_0.05- HighF_0.50	LowF_0.05- HighF_0.50	LowF_0.05- HighF_0.50
Fine	Between	Between	Between	Between
Probability	Between	Between	Between	Between
Observations	1,360	1,360	1,360	1,352
R^2	0.039	0.052	0.062	0.069

Notes: OLS regressions with robust standard errors in parentheses. The dependent variable is dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable that is coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

In Appendix C, we report the results of the treatments in which participants faced unknown fines and/or probabilities. The results show that when the detection probability was unknown, participants reacted to fines similarly to the treatments in which the probabilities were known. Furthermore, we find that making the penalty ambiguous (with known or unknown probabilities) did not further decrease lying. These results show the robustness of our findings on the deterring effects of variations in the size of fines and in the probability of being audited.

Figure 4.2 displays the fraction of lies by the expected payoff from sending a false message, assuming the Receiver followed the message. The figure shows participants were not sensitive to decreases in expected earnings that arose from an increase in the chance of being audited, but were sensitive to a decrease in expected earnings that arose from a higher fine. For any given fine level, we find no difference in lying rates across different probabilities. When the penalty for lying was low, the fraction of lies was .315 when the probability of being audited was the smallest (5%, expected payoff: $24.38), and it did not decrease when the probability was the highest (50%, expected payoff: $18.75, .342). Even when the expected payoff decreased by almost half in the high-fine scenario (from $23.75 to $12.50), the fraction of lies did not change (.246 vs. .230). We only observed a decrease in lying when the decrease in expected payoff arose from an increase in the size of the fine. Note this result has a one-to-one relation to Results 2 and 3 above.

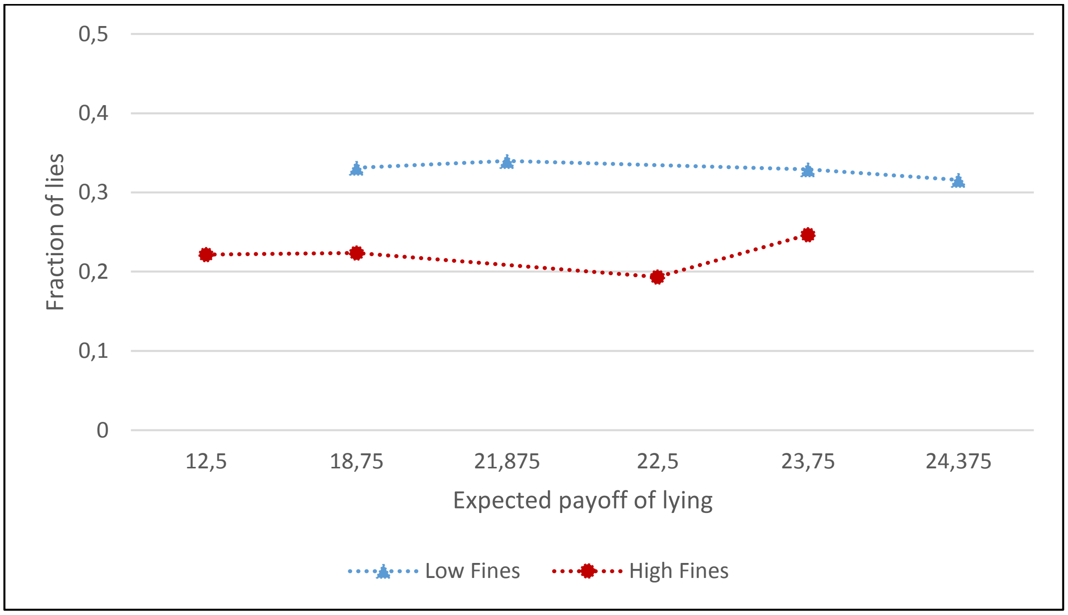


FIGURE 4.2: FRACTION OF LIES BY EXPECTED PAYOFF FROM LYING (BETWEEN-PARTICIPANT TREATMENTS)

As mentioned earlier, in two cases, the expected value from lying was identical but was determined by different combinations of fines and probabilities. In both treatment HighF_0.05 and LowF_0.10, the expected payoff was $23.75. In HighF_0.05, the probability of being audited was 5% with a high fine, whereas in LowF_0.10, this probability was 10% with a low fine. We observe a smaller fraction of lies when the fine was high (.247 and .331 in treatments HighF_0.05 and LowF_0.10, respectively), though the difference is not statistically significant (*p*=.104).

In both treatments, HighF_0.25 and LowF_0.50, the expected payoff from lying conditional on the Receiver following the message was $18.75. In HighF_0.25 (LowF_0.50), the probability of being audited was 25% (50%) when the fine was high (low). The fraction of lies was significantly smaller when the fine was high (HighF_0.25: .228 vs. LowF_0.50: .342, *p*=.029). The above findings from the equal-expected-value cases support Results 2 and 3: our participants are more sensitive to changes in the fine than to changes in the probability of being audited.

Our results are not in line with the predictions of expected utility. Non-expected utility models also cannot explain these results. For example, given the parameters estimated in the literature, Cumulative Prospect Theory does not predict that individuals will be completely insensitive to an increase in probability from 0.05 to 0.5. The complete insensitivity to probabilities we document in our data suggests that participants' decision to lie were based on decision heuristics that largely depended on the size of the fine, potentially because they are more salient or easy to evaluate. This result is in line with a larger literature on bounded rationality, which shows that individuals often use cognitive shortcuts to make their decisions, rather than relying on computations of expected utility (Simon 1955, Kahneman 2003).

The finding that lying behavior does not depend on the likelihood of being audited is in contrast with the empirical findings of a large literature in experimental economics that, in the domain of tax evasion (e.g., Friedland 1982) or in other games devised to test Becker's theory (e.g., Schildberg-Hörisch and Strassmair 2012), find individuals are sensitive to variations of detection probabilities, even small ones. A closer look at this work, however, reveals that none of this research investigated reaction to probabilities in a one-shot setting using a between-participant design as we do. Instead, several experiments use within-participant designs in which individuals are confronted with several probability parameters at the same time, which they can compare and contrast. In addition, many other experiments also investigate this question in repeated settings, where individuals can learn how to react to probabilities, especially when presented with different parameters. We review this research in Table 4.6 in Appendix A. We conjecture that in such settings, incorporating detection probabilities is much easier than doing so in settings where participants only face one given chance of being audited, presented in isolation.

In the remainder of the paper, we examine whether the insensitivity to probabilities that we detect in the one-shot and between-participant setting vanishes when individuals can directly compare different probabilities (within-participant design, section 4.4.2) or when they experience the same probability parameter over time (repeated setting, section 5).

4.4.2 Varying Fines and Probabilities: Within-Participant Data

In this section, we present the results of treatments LowF_pWithin and HighF_pWithin, where we investigate the effects of fines and probabilities using a within-participant design.

Fines between Participants and Probabilities within Participants—As Figure 4.3 shows, participants in these treatments were sensitive to the size of the fine. On average, the fraction of lies was .318 in treatment LowF_pWithin, where the fine was low, and decreased to .117 in treatment HighF_pWithin, where the fine was high (χ^2=45.48, p<.001, chi-squared test). As in the data where all parameters were presented between participants (LowF_0.05 to HighF_0.50), our results show that, conditional on a given detection probability, the difference in the fraction of lies between high and low fines is statistically significant (p<.001 for all pairwise comparisons, chi-squared test). This finding shows the robustness of our results with respect to fines, because in these treatments the fine parameters were again presented between participants.

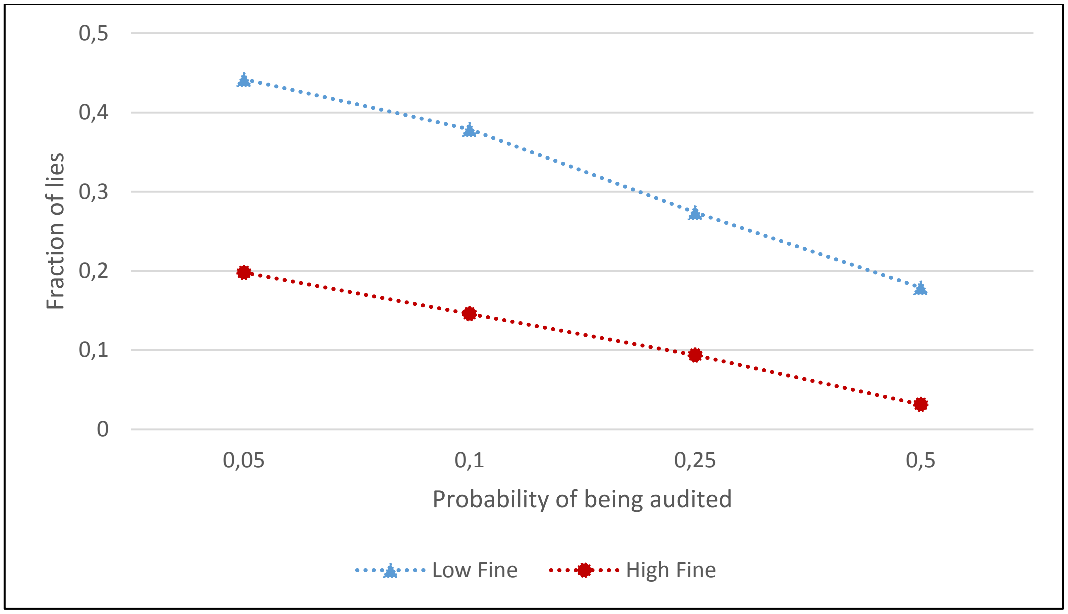


FIGURE 4.3: FRACTION OF LIES BY PROBABILITY OF BEING AUDITED AND BY FINE SIZE (WITHIN-PROBABILITY TREATMENTS)

Unlike in the between-participant data, individuals reacted to probabilities in both treatments when probability parameters were presented within participants. Lying rates were the highest when the probability of being audited was the smallest, and subsequently decreased when this probability increased. When the fine was low, the lying fraction was .442 in the former case, and it decreased to .170 in the latter (p<.001). When the fine was high, this fraction was .198 in the former case, and it decreased to .031 in the latter (p<.001).

Result 4 *When detection probabilities are presented within participants, individuals*

a. *deceive less when the fine increases, and*

b. *deceive less when the probability of detection increases.*

A different way to look at this result is to consider how participants responded to increases in expected earnings. We observe that, as the model predicts, increases in the expected value corresponded to increases in the propensity to deceive; see Figure 4.4.

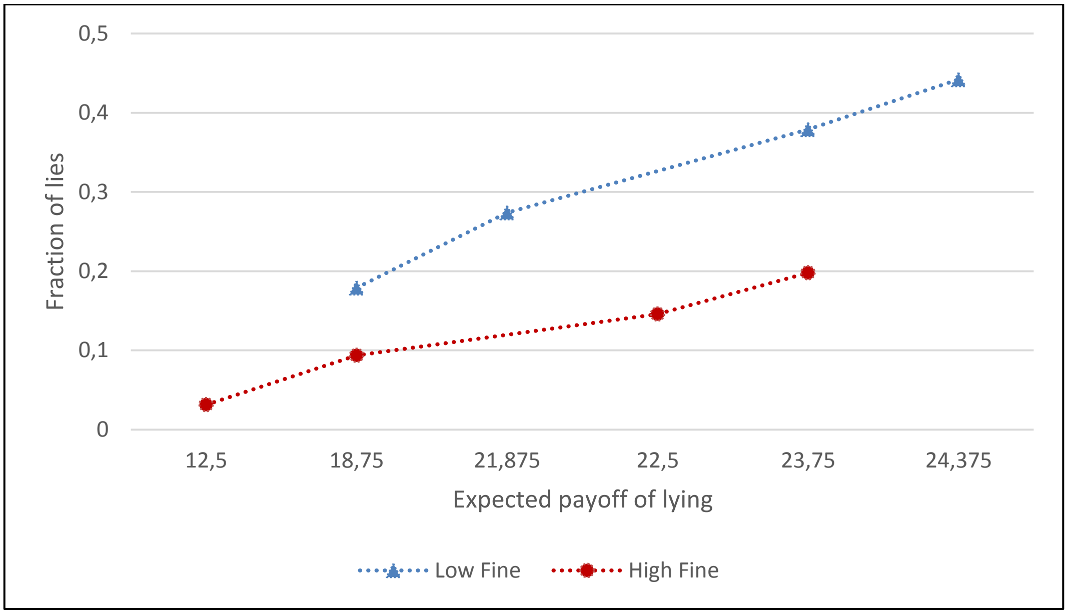


FIGURE 4.4: FRACTION OF LIES BY EXPECTED PAYOFF IF LIED (WITHIN-PROBABILITY TREATMENTS)

When comparing the cases in which the expected payoffs from lying were the same but were determined by different combinations of fines and probabilities, we still find that for a given expected payoff, participants' behavior was more strongly affected by fines. When the expected payoff was $23.75 and the fine was low, the fraction of lies was .379, whereas it was .198 when the fine was high (χ^2=7.51, *p*=.006). When the expected payoff was $18.75, the fraction of participants who lied when the fine was low was .179, whereas it was only .094 when the fine was high (χ^2=2.89, *p*=.089). This result provides further evidence that the decision to lie is more strongly affected by the size of fines than by the magnitude of detection probabilities. OLS regressions, investigating how the probability of lying changes as a function of fines, probabilities, and their interaction confirm the results (see Table 4.13 in Appendix D).

Overall, these results provide further support for the effectiveness of increasing the size of fines as a way to deter unethical behavior. In addition, they show that lying behavior is affected by increases in the detection probabilities when different probabilities are evaluated jointly. These results are consistent with experimental studies that, using within-participant designs, find individuals respond to detection probabilities (e.g., Webley 1987, Alm, McClelland, and Schulze 1992, Frank and Schulze 2000, Friesen 2012, Rizzolli and Stanca 2012, Bott 2016, see the methodological discussion in Charness, Gneezy, and Kuhn 2012).

Fixed probability and varying fine within participants—In treatments Fwithin_0.05 and Fwithin_0.50, we varied the probability of being audited between participants and the size of the fine within participants. That is, for a given probability of being audited (either 0.05, 0.10, .025, or 0.50), participants decided whether to lie in both a high-fine and a low-fine scenario. Figure 4.5 depicts the results. As in all the other treatments, we find that individuals' sensitivity to fines is robust when parameters are introduced within participants. Further, similarly to treatments LowF_0.05 to HighF_0.50 and differently from LowF_pWithin and HighF_pWithin, lying behavior was insensitive to probabilities. This finding clearly shows that the insensitivity to probabilities when individuals evaluate them in isolation is a robust finding. When the penalty was low, the fraction of participants who lied was .324 when the chance of being audited was 0.05, .317 when the chance was 0.10, .343 when the chance was 0.25, and .42 when the chance was the highest (0.50). When the fine was high, we observe a similar pattern, with similar lying rates when the probability was 0.05 and when it was 0.50 (.194).

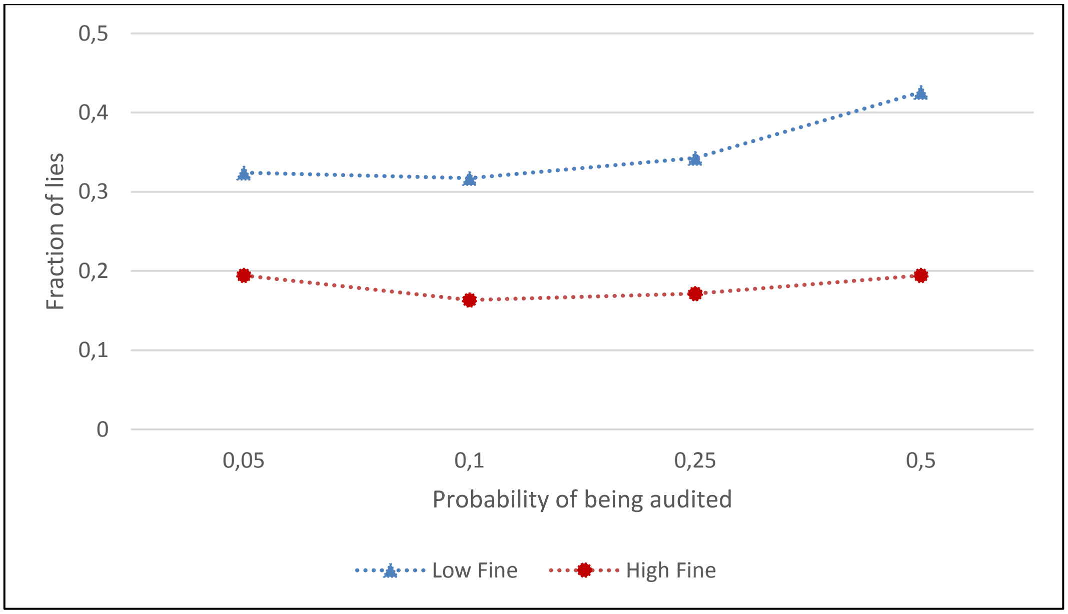


FIGURE 4.5: FRACTION OF LIES BY PROBABILITY OF BEING AUDITED AND BY FINE SIZE (WITHIN-FINE TREATMENTS)

Result 5 *When evaluating fines within participants and probabilities between participants, individuals*

a. deceive less when the fine increases, and

b. do not react to increases in the probability of being audited.

As above, when the expected payoff was the same, participants were responsive to fines and not to probabilities. Both when the expected payoff was $23.75 and when it was $18.75, participants lied significantly less often when the fine was high (.194 vs. .317, χ^2=4.19, *p*=.041 in the former case; .171 vs. .250, χ^2=16.25, p<.001 in the latter case). Table 4.14 in Appendix D explores the results using OLS regression.

4.5 Decisions from Experience

In the experiments reported above, decisions were based on a one-shot decision, with individuals relying on a description of probabilities. Our surprising finding that participants were not sensitive to changes in probabilities might be a result of them having no experience in the task. Would participants learn to react to detection probabilities over time, in the presence of immediate feedback on their behavior?

An emerging literature in psychology has documented a difference in decision-making when individuals make decisions relying on *descriptions* of probabilities and outcomes, as opposed to decisions based on *experiencing* probabilities and outcomes by directly receiving feedback on the consequences of their choices (see, e.g., Barron and Erev 2003, Hertwig, Barron, Weber, and Erev 2004, Hertwig and Erev 2009). Traditionally, many of the studies on decision under uncertainty have examined situations in which responders make only one choice per problem and rarely receive feedback on their decisions. However, the above literature argues that in the real world, people often rely on their personal experiences when making choices, rather than on a description of probabilities. This work typically compares risky choices in settings in which probabilities are described to individuals with choices in which identical probabilities are sampled over time. The findings in the literature on the description-experience gap suggest individuals tend to overweight the probability of rare events when relying on one-shot descriptions, but are less likely to do so and more likely to underweight them when making feedback-based decisions.

Based on this literature, we hypothesize that lying behavior may become more sensitive to detection probabilities in situations in which individuals have a chance to experience the auditing process and get experiential information (i.e., feedback) on the likelihood of being audited. By experiencing being audited, individuals may start putting less weight on a 5% chance of being audited, because they rarely actually experience being audited. Conversely, individuals may put more weight on the 50% chance of being audited when they experience it frequently, and thus lie less in those treatments.

We investigate this hypothesis in an additional experiment in which participants make repeated decisions with feedback after each stage. Importantly, in our design, participants receive both sources of information: the description of probabilities and the experience of auditing, as in Lejarraga and Gonzalez (2011) and Hagmann, Harman, and Gonzalez (2015).

We predicted that lying rates would increase for the 5% probability treatments, because participants would rarely experience being auditing, and would decrease for the 50% probability treatments, because participants would experience being audited frequently.

4.5.1 Experimental Design

We used a similar design as in the one-shot game, but asked participants to repeat the decision for 20 rounds, with feedback after each one. Participants started with a $10 participation fee. At the beginning of the experiment, they were presented with a description of the probability of being audited as well as of the size of the fine. In every round, the procedure for Senders was the same as in the one-shot game. Senders were informed that at the end of the experiment, the message of a randomly selected round would be sent to the Receiver, and her choice would determine the payoffs for both players, with the same payoffs as in the one-shot design.

At the end of each round, individuals received feedback on whether they were audited. If audited, participants who lied lost half (low-fine treatments) or all (high-fine treatments) of their earnings if that round was selected for payment. 49 participants took part in the experiment in the role of Receivers. Of these, 76% followed the message.[7] Importantly, in this context, feedback does not provide instrumental information, given the detection probability was clearly stated at the beginning of the experiment. Providing participants with immediate feedback on whether they were audited may facilitate the incorporation of detection probabilities in their decisions.

Participants were randomly assigned to five treatments. In Baseline, Senders were not audited. This treatment allowed us to see how, in the absence of auditing, the decision to lie evolves over

[7] 73% of the men followed the message, whereas 83% of the women did. This difference is not significant (χ^2=0.64, p<.425).

time. In four additional treatments, we introduced auditing and varied the size of the fine and the probability of being detected. As in the one-shot game, we varied the size of the fine to be either high or low. Further, we varied the probability of being detected to be either 0.05 or 0.50, focusing on the two extreme probabilities we explored in the first set of experiments. The resulting treatments are displayed in Table 4.5.

TABLE 4.5: DESCRIPTIVE STATISTICS OF LYING RATES –TREATMENTS DECISIONS FROM EXPERIENCE

	Low Fine			**High Fine**		
Probability	**Treatment**	**Mean**	**N**	**Treatment**	**Mean**	**N**
<u>Round 1</u>						
No auditing	Baseline	.397	146			
0.05	LowF_0.05	.329	213	HighF_0.05	.167	198
0.50	LowF_0.50	.311	190	HighF_0.50	.192	182
<u>Round 2-20</u>						
No auditing	Baseline	.413	146			
0.05	LowF_0.05	.297	213	HighF_0.05	.163	198
0.50	LowF_0.50	.194	190	HighF_0.50	.119	182
Overall	LowF_0.05, LowF_0.50	.332	605	HighF_0.05, HighF_0.50	.225	604

4.5.2 Results

Overall, 929 (60.6% females) participants took part in the experiment. On average, participants were 33.54 years old (SD=12.84, Median=31). In terms of ethnicity, 73.6% of the participants reported being Caucasian, 8.8% reported being African-American, 6.0% reported being Asian, 4.5% reported being Hispanic, and the remaining participants either indicated a mixed ethnicity or did not report their ethnicity. The wide majority of participants (97.81%) indicated English to be their first language. We asked the remaining participants to report their English proficiency level on a scale from 1 to 7; the average score was 6.8 (SD=.51).

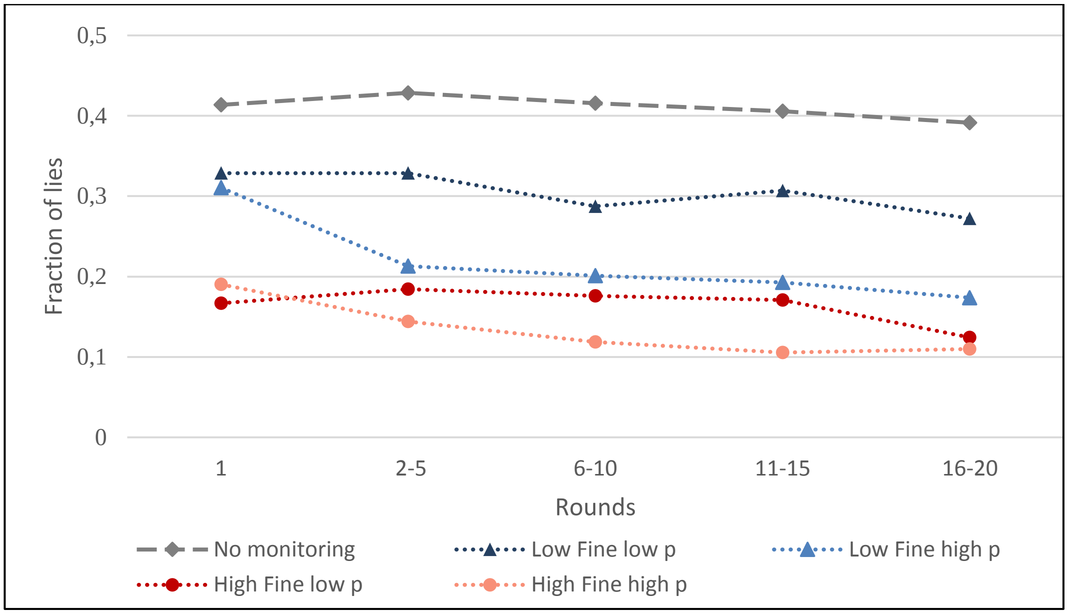


FIGURE 4.6: LYING OVER TIME – FRACTION OF LIES BY ROUNDS

Figure 4.6 displays the fraction of lies over time for all treatments. The graph shows the fraction of lies in the baseline was higher than in all the other treatments in which participants were audited, starting with round 1; see the round 1 statistics in Table 4.5.[8] When focusing on the first round, the results of the four treatments in which participants were audited replicate the ones we observed in the one-shot experiment. Lying behavior was sensitive to the size of the fine. As in the one-shot experiment, participants were not sensitive to the detection probabilities. However, starting in the second round, participants began to incorporate detection probabilities in the decision to lie. Both when the fine was high and when the fine was low, the fraction of lies dropped in the treatments with high detection probabilities. Overall, in the low-fine treatments, the average lying rate across rounds 2-20 was .297 when the detection probability was 0.05, and dropped to .194 when the detection probability was 0.50 (χ^2=108.50, p<.001).

In the high-fine treatments, the average lying rate across rounds 2-20 was .163 when the detection probability was 0.05, and dropped to .114 when the detection probability was 0.50 (χ^2=26.33, p<.001).

[8] Figure 4.9 in Appendix E shows a round-by-round graph.

Result 6 *In a repeated game with feedback, after round 1, participants are sensitive to the probability of being audited.*

In contrast to what we conjectured, lying rates in the low-probability treatments did not increase once participants experienced the low probability of being audited. Overall, the fraction of lies in these treatments was similar to the fraction of lies observed in the one-shot game. As hypothesized, participants lied less in the high-detection treatments. The effect was more pronounced for the low fine treatment, most likely because lying rates in the high-fine treatments were already very low.

Result 7 *Experience*

a. reduces deception when the detection probability is high.
b. does not affect deception when the detection probability is low.

We confirm further explore these results using regression analysis in Table 4.15 in Appendix D.

4.6 Discussion

What "works" in deterring unethical behavior? A large literature is devoted to trying to back up the effectiveness of detecting and punishing of major crimes, using field data on, for example, increases in the presence of police force. In this paper, we consider small-scale unethical behavior, and use experiments as a tool to better identify causality. In our one-shot between-participant setup, for a given expected value of penalty, increasing the size of the fine was effective in reducing unethical behavior. Instead, participants were insensitive to changes in probabilities of being audited. By contrast, in the within-participant design in which participants were able to directly compare different probability levels, increases in the detection probability did affect behavior. This sensitivity to detection probabilities in the (less realistic) within-participant design implies participants knew they should react to such changes. So why did they not react to them in the between-participant design?

One potential reason is that when evaluating detection probabilities in isolation and without experience, individuals have difficulty assessing how an objective likelihood parameter should affect their decision. Whereas behavior is responsive to the presence versus absence of monitoring, and individuals can size the magnitude of fines, they have a hard time quantifying whether a given detection probability is small or large. As a consequence, conditional on the presence of monitoring, individuals' perceived chance of being audited does not seem to change with variations in the actual chance of being audited. Therefore, individuals end up making the same decision under a wide range of probability levels, and seem to follow decision heuristics based on the size of fines.

This explanation implies that providing people with a reference point could potentially increase the sensitivity to probabilities. For example, instead of simply announcing a given detection probability, informing people that their chance of being audited has increased may provide them with a reference point that helps them incorporating detection likelihoods into their decision. An alternative approach could be to manipulate individuals' perceived chance of being detected directly, without presenting them with an objective likelihood parameter (see, e.g., Bott, Cappelen, Sørensen, and Tungodden 2014).

Our findings also illustrate that when we allowed participants to make a series of decisions with feedback (facing the same detection parameter multiple times) instead of only making one isolated decision, they became sensitive to the likelihood of being audited and lied less when the detection probability was high. This result suggests deterrence policies that rely on changes in the detection probability should take into account the key distinction between *description* versus *experience* of probabilities, and their differential effect on behavior. For example, such policies are likely to be effective for unethical behavior in which individuals receive frequent feedback on the outcome and likelihood of being audited, as in the case of fare evasion in public transportation (Dai, Galeotti, and Villeval forthcoming). However, they may be less effective in deterring unethical behavior when detection probabilities can only be presented via a description and feedback is rare, such as small-scale tax evasion.

Our results provide a novel contribution to the literature on deterrence, by showing individuals are largely insensitive to detection probabilities when such probabilities are presented in isolation. Although at first sight this finding is in contrast to other results in the literature, we reconcile it with previous work by showing that individuals become sensitive to detection probabilities in joint evaluation settings where probabilities can be directly compared or after having experienced a given detection likelihood over time. This result is important, because in many situations in which individuals make one-time choices under uncertainty, the lack of opportunities to make comparisons or learn from experience may affect the quality of their decision-making. Future work could further explore the insensitivity to detection probabilities that we document in this paper, and develop interventions to overcome it.

Overall, across all settings, we consistently find that for a given expected value, people are sensitive to increases in the size of the penalty. The magnitude of fines seems to be salient to individuals regardless of the nature of the decision environment: individuals seem to use it as a decision rule for determining whether to engage in unethical behavior. Based on these results, we conclude that harsher fines are likely to be a more successful means of deterrence of small-scale unethical behavior than increasing the probability of detection.

4.7 Appendix to Chapter 4

A Literature Review of Work on Deterrence

TABLE 4.6: REVIEW OF EXPERIMENTAL WORK ON DETERRENCE

	Within-participant design	**Repeated Game?**	**Probability levels**	**Fine levels**	**Comments on Design**	**Key Results**
Friedland, Maital, and Rutenberg (1978), *Journal of Public Economics*	✓	✓ 4 rounds	1 out of 15 or 5 out 15	3 times or 15 times the amount evaded.	N=15 Tax-evasion framing. Instructions tell participants to "maximize their net gains."	Participants were more sensitive to increases in fines than in increases in probability. Results are not statistically significant.
Friedland (1982), *Journal of Applied Social Psychology*	✓	✓ 16 rounds	7 out of 13 or 3 out of 13	3 or 7 times the sum of tax evaded	N=13 Instructions ask subjects to "maximize their net gains." Participants also receive imprecise information about the audit chances in some rounds.	Varying probability is effective. Fines are effective only when audit probabilities are low. Problem: participants play 16 round and the observations across all rounds are treated as independent in the analyses.
Spicer and Thomas (1982), *Journal of Economic Psychology*	✓	✓ 3 rounds	1 in 20 chance of being audited (round 1), 5 in 20 (round 2), 3 in 20 (round 3)	7 times the amount of tax evaded	N=54 18 participants get precise information about probabilities, 18 participants receive imprecise information about probabilities, and 18 participants get no information. Uses Friedland design (1982) and asks participants to "maximize their net gains."	With precise information about probabilities, participants tend to cheat less than in a case in which they get imprecise or no information. The results group observations across rounds and do not look at sensitivity to a given probability level.

Webley (1987), *Economics Letters*	✓ Fines varied between participants, probabilities within	✓ 2 rounds	One in 6 or 3 in 6.	2 or 6 times the amount evaded.	N=46 Business simulation where participants imagined making decisions in a hypothetical situation.	Probabilities affected percentage of income declared. The size of the fine had no effect.
Alm, McClelland, and Schulze (1992), *Journal of Public Economics*	✓	✓ 45 rounds	0, 0.02, 0.10	15 times the amount evaded	N=72 3 treatments that vary probabilities within subject (in random order). Tax-evasion framing.	Sensitive to probabilities. Compliance increases with increases in probabilities.
Beck, Davis, and Jung (1991), *The Accounting Review* (Experiment 1)		✓	50%, 40%, 90%	0.2 or 2 times of amount evaded.	N=14 Participants report income level and can be audited and receive a fine.	Both changes in fines and probabilities increase compliance.
Bott, (2016), *Working paper*	✓	✓ 3 rounds	0% in round 1, and either 5% or 25% in round 2	50% of amount evaded	N=96 per treatment Tax-evasion framing.	Participants are sensitive to increases in detection probabilities.
Schildberg-Hörisch and Strassmair (2012), *The Journal of Law, Economics, and Organization*	✓ (and between-participant design)	✓ 2 rounds	Varies between 0, 0.5 0.6, 0.7, and 0.8	Varies between 0 and 40 points, each point = 0.15 cents	Stealing game in which a decision maker can take money from a passive participant. Participants play over two rounds facing two different treatments.	Small/intermediate fines backfire and participants steal more the larger the incentive to do so. Large fines deter participants. Interchangeability of detection probability and fine.

Khadjavi (2015), *The Journal of Law, Economics and Organization*		✓ 10 rounds	Treatment 1 p=0.5. Treatment 2 p increases with the amount stolen	Treatment 1 increases with the amount stolen. Treatment 2 f=€2.5	N=408 across 3 treatment. Control treatment (no deterrence) plus two treatments in which the expected returns to stealing are the same, but determined by a different combination of p and f.	Deterrence works. Combinations of low probabilities/high fines and high probabilities/low fines are equally effective.
Harbaugh, Mocan, and Visser (2013), *Journal of Labor Research*	✓	✓ 13 rounds	Varies between 5% and 75%	Varies between $0.1 and $5.7. Initial endowment varies from $8, $12 and $16	N=82 high school students and N=34 college students Participants make 13 choices from a choice set containing 8 options. Each alternative choice involves different combinations of stolen money, probability, and fine.	Deterrence works. Fines hold a stronger relative effectiveness compared to probabilities. p and f are measured on different scales and without holding expected value constant.
Anderson and Stafford (2003), *Journal of Regulatory Economics*	✓ (and between-participant design)	✓ for within-subject design	Varies between 10% and 90%.	Varies from 1 to 4 times the amount taken.	N=unknown Public good game where individuals can free-ride and the expected penalty is contingent on the extent of free riding.	Compliance increases in expected punishment cost. Fines have a larger effect than probabilities.
DeAngelo and Charness (2012), *Journal of Risk and Uncertainty*	✓	✓ 30 periods	Regime #1 p=1/3 Regime #2 p=2/3 Regime #3 p=3/5 Regime #4 p=4/5	Regime #1 f=$0.90 Regime #2 f=$0.45 Regime #3 f=$0.833 Regime #4 f=$0.625	N=125 Roadway-speeding framing. Given information about 2 regimes (high fine, low p; low p, high fine) and are told that one of them is implemented with 50% chance. Round 11-onwards: participants vote for one of the 2 regimes, find out which one is implemented, and decide whether to speed. Round 21-30: they face two different regimes.	Violations decrease in the expected cost of speeding. When the costs are high, violations are lower when the probability of punishment is higher.

Friesen (2012), *Southern Economic Journal*	✓	✓ 30 rounds	Varies between 0 and 1 in increments of 0.1	f ranged from AU $4 to $20	N=139 In each round, participants receive a fixed amount of revenue and choose between “Complying” and “Not complying.” Different detection parameters in each round and feedback after every round. Before starting the experiment, risk preferences are elicited using Holt and Laury (2002).	Fines have a stronger deterrence effect than probabilities.

Notes: This review focuses on experimental papers that directly manipulate probability parameters and study their effect on decision-making. It does not claim to be a comprehensive review of the work on deterrence.

B Cumulative Prospect Theory Predictions

In Prospect Theory (Kahneman and Tversky 1979) and Cumulative Prospect Theory (CPT; Tversky and Kahneman 1992) the decision maker (DM) derives utility from gains and losses relative to a reference point rather than deriving utility from total earnings. Following standard assumptions of CPT, we take the value function that is kinked at the reference point such that losses loom larger than gains.

$$\text{V(x|r)} = \begin{cases} (x-r)^{\alpha} & \text{if } x \geq r \\ -\lambda(-(x-r))^{\alpha} & \text{if } x < r \end{cases}$$

The kink at the reference point is captured by the parameter $\lambda > 1$, where greater λ imply greater loss aversion. Furthermore, the value function is concave over gains and convex over losses, as captured by the parameter $\alpha \leq 1$.

Under CPT, the DM evaluates risky prospects using a probability-weighting function w(p) that transforms objective probabilities into decision weights π_i assigned to each possible outcome x_i. The weighting function is strictly increasing, w(0)=0 and w(1)=1, and the function is strictly differentiable on (0, 1). We follow Tversky and Kahneman (1992) and use the S-shaped functional form that overweights small probabilities and underweights large probabilities. The function is the same for gains and losses:

$$w(p) = \frac{p^{\delta}}{(p^{\delta} + (1-p)^{\delta})^{\frac{1}{\delta}}}$$

We examine the differential effect of monitoring on lying behavior using CPT for the parameters of our experiment. In the experiment, participant start with a $10 participation fee, P. If they decide to tell the truth, they earn an additional $5, x_T. If they choose to lie, they can either earn or lose money as compared to the final amount they would make by telling the truth. Hence, we assume the $15 earnings from P plus the payoff from telling the truth x_T to be the reference point r:

$$r = P + x_T$$

If a DM chooses to lie and is not audited, she will end the experiment with additional earnings x_L. Her final payoff will be $y = r + x_L$. That is, she will end the experiment in the gain domain. If, instead, a DM chooses to lie and is audited, a fee f_i is deducted from her final earnings, and she will end the experiment in the loss domain. Her final payoff will be $y = r + x_L - f_i$. The fee corresponds to either half or all of the DM's earnings from lying x_L, and either half or all of the participation fee P.

In the high-fine treatments, the fine is $f_H = r + x_L$, such that the DM ends the experiment with no earnings. Her final earnings are $y = r$ + \$10 - \$25 = r- \$15, that is, \$0. In the low-fine treatments, the fine is $f_L = \frac{1}{2} * (r + x_L)$, such that the DM ends the experiment with positive earnings that are yet below the reference point r. In particular, her final earnings are $y = r$ + \$10 - $\frac{1}{2}$ * (\$15 + \$10) = r - \$2.50, that is, \$12.50.

For our estimation, we use $\lambda = 2.25$, which is the loss aversion parameter estimated by Tversky and Kahneman (1992). We also assume a parameter $\alpha = 0.88$, given the small stakes involved in the experiment. Using different α parameters yields similar results.

To understand whether individuals are sensitive to changes in probabilities under CPT, we examine lying behavior for different detection-probability levels at different δ parameters. The goal of the analysis is to understand whether our results can be explained by a particularly pronounced S-shaped probability-weighting function, which would imply insensitivity to the detection probabilities given the parameters adopted in our experiment. In the model, DMs choose to lie whenever their utility from lying is greater than the utility from telling the truth, which we assume to be their reference point.

Our analysis shows that, for the low-fine treatments, DMs always lie regardless of the detection probability. The analysis for the high-fine treatments is displayed in Figure 4.7 below. The figure illustrates the DM's strategies, which imply their valuation of a risky prospect that comes from lying as compared to a sure prospect from telling the truth, for different p and δ levels. From the plot, we can see that for δ between 1 and 0.4, the DM prefers to lie when the detection probability is below approximately 0.20. However, the DM switches to telling the truth for probabilities of approximately 0.20 or higher. Hence, for these δ parameters, the DM always tells the truth when the detection probability is 0.25, which is the probability level we used in the experiment. The graph shows that for p=0.5, DMs tell the truth for δ>0.30. Only for δ<0.30 do individuals become insensitive to detection probabilities and lie at the same rate in all treatments. Note the median δ estimated in the literature is much larger than 0.30. For example, Kahneman and Tversky (1992) estimate a δ parameter of 0.61, Camerer and Ho (1994) estimate δ=0.56, and Wu and Gonzalez (1996) estimate an aggregate parameter of 0.71. Hence, only if the probability weighting function has a very pronounced S-shape, which is much more pronounced than what has been estimated in the literature so far, are the DM's insensitive to changes in detection probabilities that range from 0.05 to 0.50.

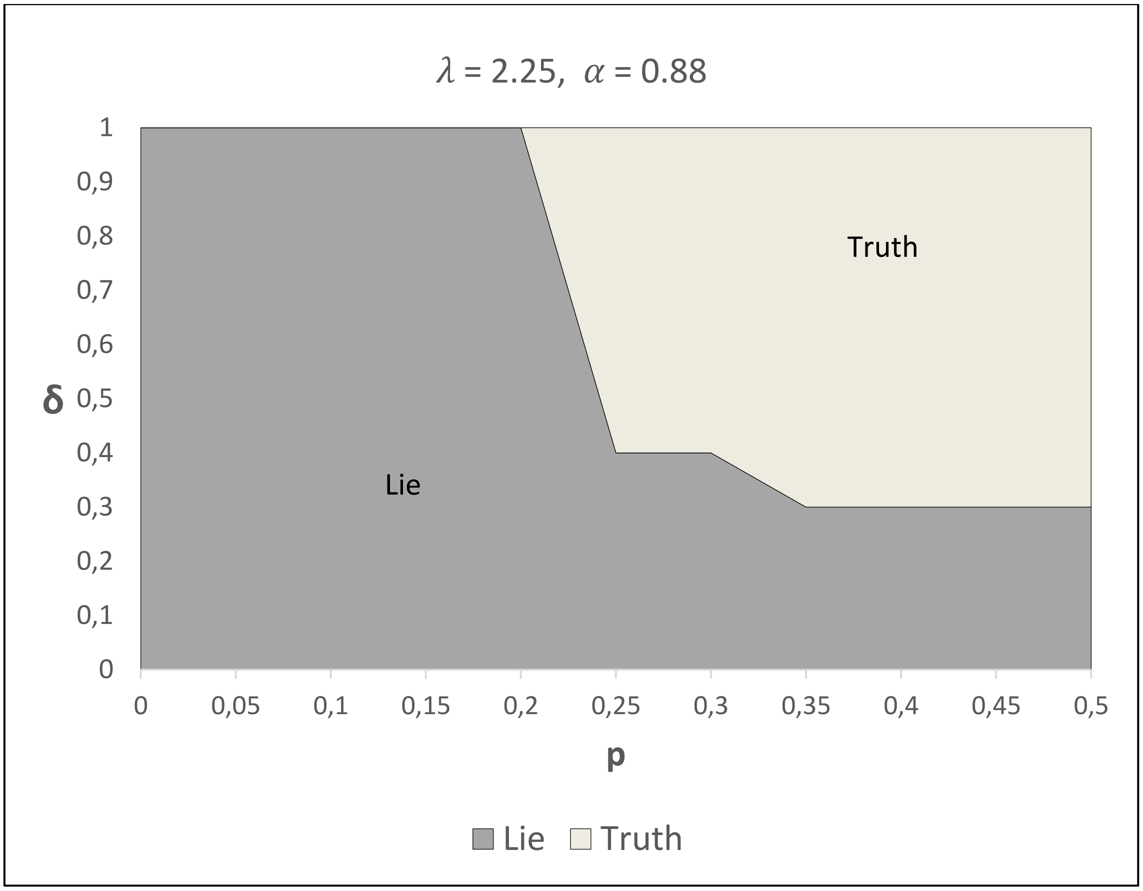


FIGURE 4.7: DECISION MAKER'S STRATEGIES BY P AND δ – HIGH-FINE TREATMENTS

C Unknown Fines and Probabilities: The Role of Ambiguity

A large body of research has demonstrated individuals are averse to ambiguous probability (Ellsberg 1961, Fox and Tversky 1995), suggesting individuals may be less likely to lie under ambiguity. We investigate the effect of introducing ambiguity about the consequences and likelihood of being audited on unethical behavior. In particular, we ran five between-participant treatments in which the size of the fine and/or the probability of being audited were ambiguous. First, we fixed the size of the fine to be either low or high (Low_pAmb and High_pAmb, respectively), and made the probability of being audited unknown. In particular, we informed the Sender that we would randomly select *some* participants and verify that their message corresponded to the actual outcome of the die roll. In these treatments, we did not specify the actual number of participants. Second, we fixed the probability of being audited to be either low (p=0.05) or high (p=0.50), and made the size of the fine ambiguous (AmbF_0.05 and AmbF_0.50). In particular, we informed the Sender that if audited, she would lose some of her earnings, without specifying how much, if the message did not correspond to the actual die roll result. Third, in AmbF_pAmb, we made both the probability of being audited and the size of the fine ambiguous. We informed the Sender that we would randomly select some participants to be audited, and, if audited, sending the wrong message would result in her losing some of her earnings. Table 4.7 below summarizes the five treatments.

TABLE 4.7: EXPERIMENTAL TREATMENTS WITH AMBIGUOUS FINES AND/OR PROBABILITIES

Treatment	**Size of the Fine** (between participants)	**Probability of being audited** (between participants)	**N**
Low_pAmb	Low	Ambiguous	156
High_pAmb	High	Ambiguous	150
AmbF_0.05	Ambiguous	0.05	151
AmbF_0.50	Ambiguous	0.50	146
AmbF_pAmb	Ambiguous	Ambiguous	152
		Total	755

Figure 4.8 below depicts the results of these treatments. The results of treatment Low_pAmb and High_pAmb show that when the detection probability was unknown, participants reacted to fines, as we show in the treatments where the probability was known. The fraction of lies in Low_pAmb was .308, whereas it was .187 in High_pAmb (chi-squared test, χ^2=6.00, *p*=.014).

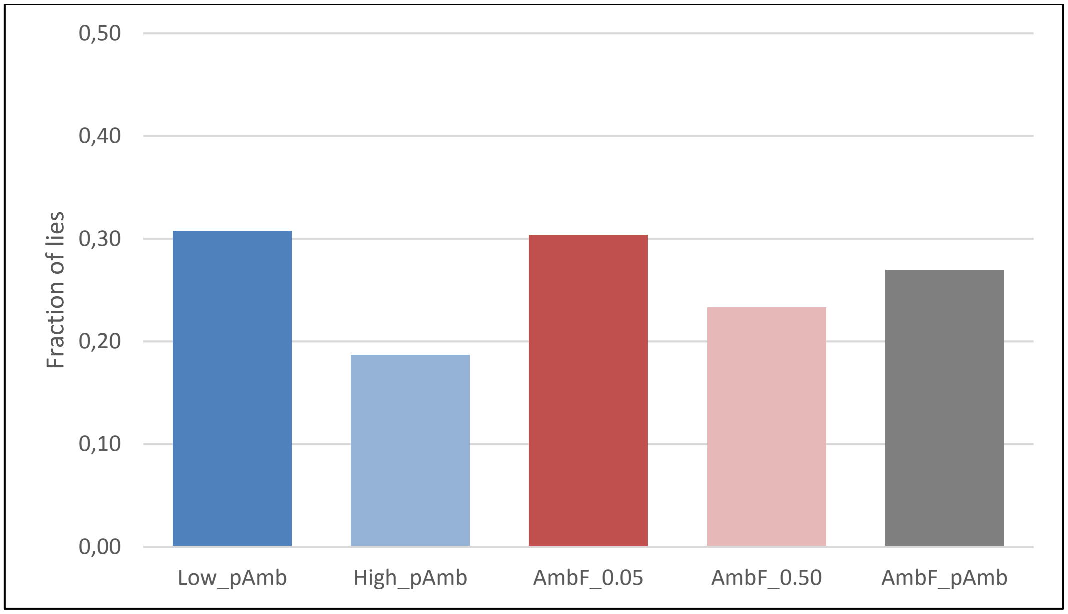


FIGURE 4.8: FRACTION OF LIES BY AMBIGUOUS PROBABILITY OF BEING AUDITED AND BY AMBIGIOUS FINE SIZE (BETWEEN-PARTICIPANT TREATMENTS)

Furthermore, we find that making the penalty ambiguous did not have a strong effect on behavior. The results of treatments AmbF_0.05 and AmbF_0.50 reveal participants lied slightly less often when the detection probability was 0.5 than when it was 0.05, but the treatment difference was not statistically significant (.30 vs .23, chi-squared test, χ^2=1.895 *p*=.169).

Finally, the results of treatment Amb_Amb show that when both the probability of being audited and the penalty were ambiguous, the fraction of participants who lied was 0.27. This fraction is larger than the fraction of lies when the probability of being audited was ambiguous and the fine was high (χ^2=2.95, *p*<.01). This fraction does not significantly differ from any of the other ambiguity treatments (all p-values are above *p*=.463). The regression reported in Table 4.8 confirms these results (see also Table 4.12 in Appendix D).

Table 4.8: Effect of Fine and Probability on Lying with Abiguity Treatments

Dependent variable:	Lied (1=Yes)			
	(1)	(2)	(3)	(4)
High Fine	-0.12***	-0.12***	-0.13***	-0.13***
	(0.031)	(0.030)	(0.030)	(0.030)
Probability	-0.04	-0.05	-0.03	-0.04
	(0.074)	(0.074)	(0.073)	(0.073)
High Fine*Probability	0.06	0.05	0.04	0.04
	(0.114)	(0.114)	(0.113)	(0.113)
Ambiguity Fine	-0.06*	-0.06**	-0.07**	-0.07**
	(0.031)	(0.031)	(0.031)	(0.031)
Ambiguity Probability	-0.04	-0.03	-0.03	-0.03
	(0.032)	(0.032)	(0.031)	(0.031)
Ambiguity Fine*Probability	0.02	0.01	0.01	0.01
	(0.055)	(0.054)	(0.054)	(0.054)
Female		-0.10***	-0.09***	-0.08***
		(0.020)	(0.020)	(0.020)
Age			-0.00***	-0.00***
			(0.001)	(0.001)
Ethnicity	No	No	No	Yes
Constant	0.34***	0.40***	0.54***	0.52***
	(0.023)	(0.025)	(0.039)	(0.040)
Treatments	Between-participant treatments	Between-participant treatments	Between-participant treatments	Between-participant treatments
Fine	Between	Between	Between	Between
Probability	Between	Between	Between	Between
Observations	1961	1961	1961	1950
R^2	.010	.022	.033	0.036

Notes: OLS regression with robust standard errors in parentheses. The dependent variable is dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half of their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

D Regression Analyses

TABLE 4.9: EFFECT OF FINE AND PROBABILITY ON LYING WITH FEMALE INTERACTION TERMS

Dependent variable:	Lied (1=Yes)	
	(1)	(2)
High Fine	-0.119*	-0.120**
	(0.061)	(0.061)
Probability	0.091	0.080
	(0.160)	(0.161)
High Fine*Probability	-0.151	-0.153
	(0.220)	(0.221)
No auditing	0.200***	0.203***
	(0.051)	(0.050)
Female	-0.130***	-0.119**
	(0.050)	(0.050)
Female*High Fine	0.039	0.039
	(0.075)	(0.074)
Female*Probability	-0.066	-0.027
	(0.194)	(0.193)
Female*High Fine*Probability	0.193	0.165
	(0.276)	(0.276)
Age		-0.004***
		(0.001)
Ethnicity	No	Yes
Constant	0.33***	0.38***
	(0.034)	(0.036)
Treatments	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50
Fine	Between	Between
Probability	Between	Between
Observations	1,360	1,352
R^2	0.054	0.071

Notes: OLS regression with robust standard errors in parentheses. The dependent variable is a dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

TABLE 4.10: EFFECT OF FINE AND PROBABILITY ON LYING – RESTRICTED SAMPLE I (ONLY THOSE WHO HAD NO DIFFICULTIES UNDERSTANDING THE TASK)

Dependent variable:	Lied (1=Yes)			
	(1)	(2)	(3)	(4)
High Fine	-0.10**	-0.10**	-0.10**	-0.10**
	(0.045)	(0.044)	(0.044)	(0.044)
Probability	0.05	0.05	0.06	0.06
	(0.119)	(0.117)	(0.116)	(0.117)
High Fine*Probability	-0.09	-0.10	-0.10	-0.12
	(0.156)	(0.154)	(0.153)	(0.155)
No auditing	0.20***	0.21***	0.21***	0.21***
	(0.054)	(0.053)	(0.053)	(0.053)
Female		-0.11***	-0.09***	-0.09***
		(0.026)	(0.026)	(0.027)
Age			-0.00***	-0.00***
			(0.001)	(0.001)
Ethnicity	No	No	No	Yes
Constant	0.33***	0.38***	0.52***	0.51***
	(0.034)	(0.036)	(0.053)	(0.054)
Treatments	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50
Fine	Between	Between	Between	Between
Probability	Between	Between	Between	Between
Observations	1,208	1,208	1,208	1,200
R^2	0.046	0.059	0.069	0.075

Notes: OLS regression with robust standard errors in parentheses Restricted to those participants who indicated in the questionnaire that they did not have any difficulties in understanding the task. The dependent variable is a dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

TABLE 4.11: EFFECT OF FINE AND PROBABILITY ON LYING – RESTRICTED SAMPLE II (ONLY THOSE WHO ANSWERED THE CONTROL QUESTION CORRECTLY)

Dependent variable:	Lied (1=Yes)			
	(1)	(2)	(3)	(4)
High Fine	-0.11**	-0.10**	-0.11**	-0.10**
	(0.048)	(0.048)	(0.048)	(0.048)
Probability	-0.00	0.01	0.02	0.02
	(0.128)	(0.126)	(0.125)	(0.125)
High Fine*Probability	-0.10	-0.12	-0.13	-0.15
	(0.166)	(0.164)	(0.164)	(0.165)
No auditing	0.23***	0.23***	0.23***	0.23***
	(0.059)	(0.059)	(0.059)	(0.059)
Female		-0.09***	-0.08***	-0.08***
		(0.028)	(0.028)	(0.028)
Age			-0.00**	-0.00**
			(0.001)	(0.001)
Ethnicity	No	No	No	Yes
Constant	0.35***	0.39***	0.49***	0.49***
	(0.037)	(0.039)	(0.057)	(0.059)
Treatments	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50
Fine	Between	Between	Between	Between
Probability	Between	Between	Between	Between
Observations	1,034	1,034	1,034	1,028
R^2	0.058	0.068	0.074	0.078

Notes: OLS regression with robust standard errors in parentheses. The dependent variable is a dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

TABLE 4.12: EFFECT OF FINE AND PROBABILITY ON LYING WITH TREATMENT DUMMIES

Dependent variable:	Lied (1=Yes)			
	(1)	(2)	(3)	(4)
LowF_0.10	0.02	0.01	0.02	0.02
	(0.052)	(0.051)	(0.051)	(0.051)
LowF_0.25	0.02	0.02	0.03	0.03
	(0.051)	(0.051)	(0.051)	(0.051)
LowF_0.50	0.03	0.03	0.03	0.03
	(0.052)	(0.052)	(0.051)	(0.051)
HighF_0.05	-0.07	-0.07	-0.06	-0.06
	(0.051)	(0.051)	(0.051)	(0.051)
HighF_0.10	-0.12**	-0.13**	-0.13***	-0.14***
	(0.052)	(0.052)	(0.051)	(0.051)
HighF_0.25	-0.09*	-0.10*	-0.10*	-0.10*
	(0.052)	(0.052)	(0.051)	(0.051)
HighF_0.50	-0.09*	-0.09*	-0.09*	-0.09*
	(0.052)	(0.051)	(0.051)	(0.051)
High_pAmb	-0.13**	-0.12**	-0.13**	-0.13**
	(0.052)	(0.051)	(0.051)	(0.051)
Low_pAmb	-0.01	-0.01	0.00	0.01
	(0.051)	(0.051)	(0.051)	(0.051)
AmbF_0.50	-0.08	-0.09*	-0.08	-0.08
	(0.052)	(0.052)	(0.052)	(0.052)
AmbF_0.05	-0.01	-0.01	-0.01	-0.01
	(0.052)	(0.052)	(0.051)	(0.051)
AmbF_pAmb	-0.05	-0.05	-0.05	-0.06
	(0.052)	(0.051)	(0.051)	(0.051)
No auditing	0.20***	0.20***	0.20***	0.21***
	(0.052)	(0.051)	(0.051)	(0.051)
Female		-0.10***	-0.08***	-0.09***
		(0.019)	(0.020)	(0.020)
Age			-0.00***	-0.00***
			(0.001)	(0.001)
Ethnicity	No	No	No	Yes
Constant	0.32***	0.37***	0.52***	0.50***
	(0.037)	(0.038)	(0.047)	(0.048)
Treatments	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50	LowF_0.05-HighF_0.50
Fine	Between	Between	Between	Between
Probability	Between	Between	Between	Between
Observations	2,112	2,112	2,112	2,100
R^2	0.032	0.044	0.056	0.056

Notes: OLS regression with robust standard errors reported in parentheses. The dependent variable is a dummy coded as 1 if a participant lied, and zero otherwise. HighF_0.05 to LowF_0.50 are dummies indicating the treatments. Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

TABLE 4.13: EFFECT OF FINE AND PROBABILITY ON LYING – WITHIN-PROBABILITY TREATMENTS

Dependent variable:	Lied (1=Yes)			
	(1)	(2)	(3)	(4)
High Fine	-0.25***	-0.25***	-0.25***	-0.26***
	(0.067)	(0.067)	(0.066)	(0.066)
Probability	-0.56***	-0.56***	-0.56***	-0.56***
	(0.113)	(0.113)	(0.114)	(0.114)
High Fine*Probability	0.22	0.22	0.22	0.22
	(0.146)	(0.146)	(0.146)	(0.146)
Female		-0.07	-0.06	-0.05
		(0.046)	(0.046)	(0.046)
Age			-0.00	-0.00
			(0.002)	(0.002)
Ethnicity	No	No	No	Yes
Constant	0.44***	0.48***	0.57***	0.55***
	(0.053)	(0.061)	(0.093)	(0.094)
Treatments	LowF_pWithin, HighF_pWithin	LowF_pWithin, HighF_pWithin	LowF_pWithin, HighF_pWithin	LowF_pWithin, HighF_pWithin
Fine	Between	Between	Between	Between
Probability	Within	Within	Within	Within
Observations	764	764	764	764
Clusters	191	191	191	191
R^2	0.098	0.105	0.110	0.130

Notes: OLS regression with robust standard errors clustered on individual level in parentheses. The dependent variable is a dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable that is coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. Within refers to within-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

We investigate how the probability of lying changes as a function of fines, probabilities, and their interaction. Standard errors are clustered at the individual level. Column 1 confirms that participants' tendency to lie decreased both with an increase in the size of the fine and with increases in the probability of being audited. We find no significant interaction, which suggests a given probability of being audited did not affect behavior differently depending on the size of the fine. Columns 2-4 confirm this result is robust to demographic controls. Unlike the data in which all parameters were presented between participants, we find no gender and age effects in these treatments.

TABLE 4.14: EFFECT OF FINE PROBABILITY OF LYING – WITHIN-FINE TREATMENTS

Dependent variable:	Lied (1=Yes)			
	(1)	(2)	(3)	(4)
High Fine	-0.12***	-0.12***	-0.12***	-0.13***
	(0.034)	(0.034)	(0.034)	(0.035)
Probability	0.24*	0.24*	0.25*	0.24*
	(0.134)	(0.134)	(0.133)	(0.134)
High Fine*Probability	-0.21	-0.21	-0.21	-0.21
	(0.131)	(0.131)	(0.131)	(0.132)
Female		-0.04	-0.02	-0.03
		(0.037)	(0.036)	(0.037)
Age			-0.00**	-0.00**
			(0.002)	(0.002)
Ethnicity	No	No	No	Yes
Constant	0.30***	0.32***	0.44***	0.45***
	(0.037)	(0.042)	(0.070)	(0.073)
Treatments	Fwithin_0.05-Fwithin_0.50	Fwithin_0.05-Fwithin_0.50	Fwithin_0.05-Fwithin_0.50	Fwithin_0.05-Fwithin_0.50
Fine	Within	Within	Within	Within
Probability	Between	Between	Between	Between
Observations	850	850	848	844
Clusters	425	425	424	422
R^2	.042	.044	.051	.056

Notes: OLS regressions with robust standard errors clustered on individual level in parentheses. The dependent variable is dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. Within refers to within-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

As column 1 shows, participants were sensitive to increases in the size of the fine. Unexpectedly, they lied more rather than less often when the probability of being audited increased, though the effect was only marginally significant and driven by a larger fraction of participants lying when the probability was 0.50. The effect of fine and probabilities is robust to demographic controls (columns 2-4). In these data, we find no gender differences in the propensity to lie.

Table 4.15: Lying over time - Effect of Fine and Probabilityy on Lying

Dependent variable:	Lied (1=Yes)					
	(1)	(2)	(3)	(4)	(5)	(6)
High Fine	-0.16***	-0.14***	-0.13***	-0.13***	-0.13***	-0.14***
	(0.043)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
High Probability	-0.02	-0.10***	-0.10***	-0.09***	-0.07**	-0.08**
	(0.044)	(0.032)	(0.032)	(0.033)	(0.035)	(0.035)
High Fine*Probability	0.04	0.06	0.06	0.06	0.06	0.07
	(0.063)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)
No auditing	0.07	0.11***	0.11***	0.12***	0.12***	0.11***
	(0.047)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
Audited in previous round			-0.01**	0.03**	0.03**	0.03**
			(0.006)	(0.015)	(0.015)	(0.016)
Audited in previous round*High Probability				-0.06***	-0.06***	-0.06***
				(0.016)	(0.016)	(0.017)
Audited in round 1					-0.05*	-0.05*
					(0.025)	(0.024)
Number of times audited					0.00	0.00
					(0.002)	(0.002)
Female						-0.06***
						(0.022)
Age						-0.00
						(0.001)
Ethnicity	No	No	No	No	No	Yes
Round Dummy	No	Yes	Yes	Yes	Yes	Yes
Constant	0.33***	0.34***	0.33***	0.33***	0.33***	0.33***
	(0.030)	(0.026)	(0.027)	(0.027)	(0.027)	(0.027)
Round	First Round	All Rounds	Rounds 2-20	Rounds 2-20	Rounds 2-20	Rounds 2-20
Fine	Between	Between	Between	Between	Between	Between
Probability	Between	Between	Between	Between	Between	Between
Observations	929	18,580	17,651	17,651	17,651	17,651
Clusters		929	929	929	929	929

Notes: Column 1 reports OLS regression with robust standard errors reported in parentheses. Columns 2-6 report estimates from a random-effects regression model with robust standard errors clustered on individual level in parentheses. The dependent variable is dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50). Audited in previous round is a lagged variable that indicates whether a given individual was audited in the previous round. Fined in previous round is a lagged variable indicating whether a given individual was fined in the previous round. Number of times audited is a continuous variable that indicates the total number of times a participant was audited. Female is a dummy variable coded as 1 if the participant was a woman, and zero otherwise. Age is a continuous variable indicating participants' age. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

In Table 4.15, we regress lying on a dummy variable coded as 1 when the fine was high and zero otherwise, a dummy variable coded as 1 when the detection probability was high (50%) and zero otherwise, and their interaction. The model also includes a dummy variable coded as 1 in the absence of auditing. Column 1 reports the results of an OLS regression that examines the data of round 1 only. As in the one-shot experiment, we observe that lying behavior is sensitive to the size of the fine but does not respond to a large change in the detection probability (5% vs. 50%).

In the remaining columns, we estimate a random-effects regression model using the data of all rounds and clustering the standard error at the individual level. The results in column 2 confirm that in presence of auditing, increasing the size of the fine from low to high decreased the likelihood of lying by 14 percentage points. However, in contrast to the analysis of the first round only, we see individuals also responded to increases in the detection probability. In particular, subjects were 10 percentage points less likely to lie when the detection probability was high. The coefficient of the interaction term is small and not significant, suggesting the effectiveness of a high fine did not depend on the detection probability. Conversely, we observe that in the absence of auditing (control treatment), participants were 11 percentage points more likely to lie. In column 3, we include a lagged variable to take into account how being audited in a previous round affected lying behavior in the subsequent round. The coefficient shows that being audited in the previous round decreased the lying probability by 1 percentage point. However, allowing this coefficient to vary according to the detection probability by interacting this variable with the high probability dummy (column 4) shows that when the detection probability was 5%, being audited in the previous round slightly increased subsequent lying behavior. By contrast, when the detection probability was high, being audited in the previous round significantly decreased lying. These results imply that unlike in the high-probability treatments, being audited in the low-probability treatments did not deter but rather increased subsequent lying, perhaps because individuals anticipate that being caught was a rare event. Importantly, when including these variables in the model, the coefficient of the high-probability dummy does not change, suggesting the effect was not driven only by individuals who were audited. In column 5, we add control variables for whether participants were audited in round 1 and for the number of times they were audited in all previous rounds. We find that being audited in round 1 decreased lying, though all other effects stay significant. These effects are robust to demographic controls (column 6). As in the one-shot game, in this experiment, women lied less often than men. In Table 4.16 we show the regression analysis with treatment dummies and in Table 4.17 we show results for the round-by-round analysis.

Table 4.16: Lying Over Time – Effect of Fine and Probabilityy on Lying with Treatment Dummies

Dependent variable:	Lied (1=Yes)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HighF_0.05	-0.16***	-0.14***	-0.13***	-0.13***	-0.13***	-0.13***	-0.14***
	(0.042)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
HighF_0.50	-0.14***	-0.18***	-0.17***	-0.17***	-0.17***	-0.15***	-0.16***
	(0.044)	(0.031)	(0.031)	(0.032)	(0.032)	(0.036)	(0.035)
LowF_0.50	-0.02	-0.10***	-0.10***	-0.08**	-0.08**	-0.07*	-0.08**
	(0.047)	(0.032)	(0.032)	(0.033)	(0.033)	(0.036)	(0.035)
No auditing	0.07	0.11***	0.11***	0.12***	0.12***	0.12***	0.11***
	(0.052)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
Audited in previous round			-0.01**	0.05**	0.05**	0.05**	0.05**
			(0.006)	(0.022)	(0.024)	(0.024)	(0.024)
Audited in previous round*High Probability				-0.09***	-0.09***	-0.09***	-0.09***
				(0.024)	(0.026)	(0.026)	(0.027)
Audited in previous round*High Fine				-0.03	-0.03	-0.03	-0.03
				(0.030)	(0.031)	(0.031)	(0.032)
Audited in previous round*High Fine*High Probability				0.05	0.06	0.06	0.06
				(0.033)	(0.034)	(0.034)	(0.035)
Audited*Fined in previous round					0.01	0.01	0.00
					(0.045)	(0.045)	(0.046)
Audited*Fined in previous round*High Probability					0.05	0.05	0.05
					(0.052)	0.052)	(0.053)
Audited in round 1						-0.05*	-0.05*
						(0.024)	(0.024)
Number of times audited						0.00	0.00
						(0.002)	(0.002)
Female							-0.06***
							(0.022)
Age							-0.00
							(0.001)
Ethnicity	No	No	No	No	No	No	Yes
Round Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	0.33***	0.34***	0.33***	0.33***	0.33***	0.33***	0.36***
	(0.032)	(0.026)	(0.027)	(0.027)	(0.027)	(0.027)	(0.046)
Round	First Round	All Rounds	All Rounds	All Rounds	All Rounds	All Rounds	All Rounds
Fine	Between	Between	Between	Between	Between	Between	Between
Probability	Between	Between	Between	Between	Between	Between	Between
Observations	929	18,580	17,651	17,651	17,651	17,651	17,290
Clusters		929	929	929	929	929	910

Notes: OLS regression with robust standard errors clustered on individual in parentheses. The dependent variable is a dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Audited in previous round is a lagged variable that indicates whether a given individual was audited in the previous round. Fined in previous round is a lagged variable indicating whether a given individual was fined in the previous round. Ethnicity differentiates between White, Black, Asian, Hispanic and mixed race. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

In Table 4.16 we investigate lying over time using dummy variables for each treatment. We see that in round one, as in the one-shot game and the analysis reported in Table 4.6 in the main text, participants in the high fine treatments lied significantly less than those in the low fine treatments. Higher detection probabilities did not deter lying in either treatment. When the fine was low, lying behavior only decreased by 2 percentage points (*p*=.698) when the detection probability was high; when the fine was high, lying increased by 2 percentage points (*p*=.517) when the detection probability was high. To investigate lying behavior across all rounds, columns 2-7 present the results of a random-effects regression model with standard errors clustered at the individual level. These regressions show participants became sensitive to probabilities over time. For example, Column 2 shows that in the low-fine treatment, lying decreased by 10 percentage points (*p*=.002) when the probability moved from low (0.05) to high (0.5). When the fine was high, lying behavior in the high-probability treatment decreased by 4 percentage points, though this decrease did not reach statistical significance (*p*=.153), possibly due to a floor effect.

TABLE 4.17: LYING OVER TIME – EFFECT OF FINE AND PROBABILITYY ON LYING BY ROUND

Dependent variable:	Lied (1=Yes)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6-10	Round 11-15	Round 16-20
High Fine	-0.16***	-0.13***	-0.17***	-0.12***	-0.17***	-0.11***	-0.14***	-0.15***
	(0.042)	(0.043)	(0.043)	(0.041)	(0.043)	(0.035)	(0.035)	(0.033)
High Probability	-0.02	-0.08*	-0.13***	-0.11***	-0.14***	-0.09**	-0.11***	-0.10***
	(0.047)	(0.045)	(0.045)	(0.041)	(0.044)	(0.034)	(0.035)	(0.034)
High Fine*Probability	0.04	0.04	0.09	0.08	0.10*	0.03	0.05	0.09*
	(0.061)	(0.060)	(0.059)	(0.055)	(0.059)	(0.046)	(0.046)	(0.044)
No auditing	0.07	0.11**	0.09	0.14***	0.07	0.13***	0.10**	0.12***
	(0.052)	(0.052)	(0.053)	(0.051)	(0.053)	(0.043)	(0.044)	(0.043)
Constant	0.33***	0.32***	0.35***	0.28***	0.36***	0.29***	0.31***	0.27***
	(0.032)	(0.032)	(0.033)	(0.031)	(0.033)	(0.026)	(0.026)	(0.026)
Fine	Between	Between	Between	Between	Between	Between	Between	Between
Probability	Between	Between	Between	Between	Between	Between	Between	Between
Observations	929	929	929	929	929	4,645	4,645	4,645
Clusters	-	-	-	-	-	929	929	929
Adj. R^2	.439	.433	.430	.408	.428			

Notes: Columns 1-5 report OLS regressions with robust standard errors reported in parentheses for the first five rounds separately. Data are clustered at the individual level. Columns 6 to 8 report estimates of a random-effects regression model for rounds 6-10, 11-15, and 16-20, respectively with robust standard errors reported in parentheses and data are clustered on individual level. The dependent variable is a dummy coded as 1 if a participant lied, and zero otherwise. High Fine is a dummy variable coded as 1 for the treatments in which participants lost all their earnings if audited, and zero for the treatments in which participants lost half their earnings if audited. Probability is a continuous variable indicating the chance of being audited (either 0.05, 0.10, 0.25, or 0.50. Between refers to between-participant data. The stars indicate significance levels: * p-value<0.1, ** p-value<0.05, *** p-value<0.01.

In Table 4.17 we report OLS estimates for the first five rounds separately, and then we report the estimates of a random-effects regression model for rounds 6-10, 11-15, and 16-20. We observe that whereas individuals were not sensitive to detection probabilities in round one (*p*=.698), they immediately became sensitive to probabilities in round 2 (*p*=.068), and the effect of probabilities slightly increased in size and became significant at the 1% level over subsequent rounds.

E Graphical Analysis

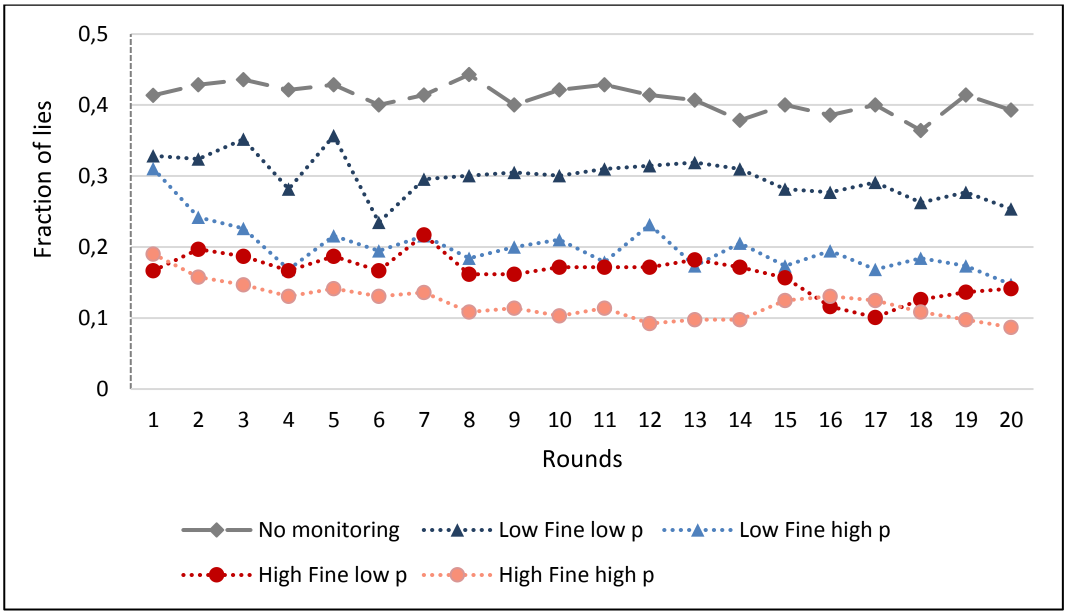


FIGURE 4.9: LYING OVER TIME – ROUND-BY-ROUND

F Instructions for Participants in the Role of Sender (Translation from German)

Welcome to our short experiment!

You will receive $0.25 for your time. On top of that, we will randomly choose 1 participant out of 20 and pay them a monetary bonus on Mturk according to the instruction that follows. Please read these instructions carefully. You may earn a considerable sum of money, depending on the decisions you make in the experiment. The rest of the instructions describe the procedure of the experiment.

----- next screen ------

If selected at random, you will receive a **$10 participation fee** and you will be matched randomly with another participant (another MTurk worker), who will receive a $0.25 participation fee. Neither of you will know the worker ID of the other.

Before starting this experiment, we have rolled a 10-sided die, and obtained the outcome 6. The other participant will not be informed that the outcome of the die roll was 6. However, he or she will be told that you have been informed about the outcome of the die roll.

Your Task

We now ask you to send a message to the other participant. The message will correspond to a number from 1 to 10. There are ten possible messages:

Message 1: “The outcome from the roll of the 10-sided die is 1.”
Message 2: “The outcome from the roll of the 10-sided die is 2.”
Message 3: “The outcome from the roll of the 10-sided die is 3.”
Message 4: “The outcome from the roll of the 10-sided die is 4.”
Message 5: “The outcome from the roll of the 10-sided die is 5.”
Message 6: “The outcome from the roll of the 10-sided die is 6.”
Message 7: “The outcome from the roll of the 10-sided die is 7.”
Message 8: “The outcome from the roll of the 10-sided die is 8.”
Message 9: “The outcome from the roll of the 10-sided die is 9.”
Message 10: “The outcome from the roll of the 10-sided die is 10.”

Your message will be shown to the other participant, and then he or she will be asked to choose a number between 1 and 10. The message you send is the only information he or she will have regarding the roll of the die.

Payments

Importantly, the number chosen by the other participant will determine the payments in the experiment, which you will receive on top of the $10 participation fee. The two possible monetary payments are:

Option A: If the other participant chooses the actual outcome of the roll of the dice, 6, then you will receive **$5** and he or she will receive **$15**.

Option B: If the other participant chooses a number different than the actual outtcome, you will receive a **$15** and he or she will receive **$5**.

The other participant will never know the actual payments associated with each message. All he or she will know is the amount of money they will be paid and that you know the payments connected to each message.

[**Treatments**: see below the variations of this message]

There is a chance that we will audit your message: We will randomly select **1 participant out of __(PROBABILITY)** and verify that the number indicated in their message corresponds to the actual outcome of the die roll.

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, you will lose both the $10 participation fee and the payment from the die roll experiment, leaving you with $0.25 earnings for this task **(FINE)**.]

Please choose the message you would like to send to the other participant.

○ "The outcome from the roll of the 10-sided die is 1."
○ "The outcome from the roll of the 10-sided die is 2."
○ "The outcome from the roll of the 10-sided die is 3."
○ "The outcome from the roll of the 10-sided die is 4."
○ "The outcome from the roll of the 10-sided die is 5."
○ "The outcome from the roll of the 10-sided die is 6."
○ "The outcome from the roll of the 10-sided die is 7."
○ "The outcome from the roll of the 10-sided die is 8."
○ "The outcome from the roll of the 10-sided die is 9."
○ "The outcome from the roll of the 10-sided die is 10."

Treatment variations.

Treatment 0: Control: regular deception game (without mentioning that there is a zero chance of being audited)
Treatment 1-8: We cross probability and fine (4x2)

PROBABILITY

There is a chance that we will audit your message. We will randomly select **1 participant out of __** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

1. 1 out of 20 (p=.05)
2. 1 out of 10 (p=.10)
3. 1 out of 4 (p=.25)
4. 1 out of 2 (p=0.5)

FINE

High: If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose both the $10 participation fee and the payment from the die roll experiment, leaving you with $0.25 earnings for this task.**

Low: If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose half of the $10 participation fee and half of the payment from the die roll experiment, leaving you with $12.75 earnings for this task.**

Treatment 9: Ambiguity about probability of getting a low fine

There is a chance that we will audit your message: We will randomly select **some participants** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose both the $10 participation fee and the payment from the die roll experiment, leaving you with $12.75 earnings for this task.**

Treatment 10: Ambiguity about probability of getting a high fine

There is a chance that we will audit your message: We will randomly select **some participants** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose both the $10 participation fee and the payment from the die roll experiment, leaving you with $0.25 earnings for this task.**

Treatment 11: Ambiguity about fine -Low probability

There is a chance that we will audit your message: We will randomly select **1 participant out of 20** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose some of your earnings for this task.**

Treatment 12: Ambiguity about fine -High probability

There is a chance that we will audit your message: We will randomly select **1 participant out of 2** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose some of your earnings for this task.**

Treatment 13: Ambiguity about both

There is a chance that we will audit your message: We will randomly select **some participants** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose some of your earnings for this task (FINE).**

Treatment 14: Strategy method with low fine

There is a chance that we will audit your message. We will randomly select **some participants** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

In particular we will audit 1 out of either 20 participants, 10 participants, 4 participants or 2 participants.

We now ask you to choose the message you would like to send for each of the possible chance of being audited. At the end of the experiment, you will be notified about what the chance is and the corresponding message will be implemented.

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose both the $10 participation fee and the payment from the die roll experiment, leaving you with $0.25 earnings for this task.**

Please choose the message you would like to send to the other participant for each of the possible options.

1. **1 participant out of 20**, **High fine**
2. **1 participant out of 10**, **High fine**
3. **1 participant out of 4**, **High fine**
4. **1 participant out of 2, High fine**

Treatment 15: Strategy method with high fine

There is a chance that we will audit your message. We will randomly select **some participants** and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

In particular we will audit 1 out of *X* participants, where *X* could correspond to either 20, 10, 4, or 2 participants.

We now ask you to choose the message you would like to send for each of the possible *Xs*. At the end of the experiment, you will be notified about what X is and the corresponding message will be implemented.

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose half of the $10 participation fee and half of the payment associated to with the message you sent, leaving you with $12.75 earnings for this task.**

Please choose the message you would like to send to the other participant for each of the possible options.

1. **1 participant out of 20**, **Low fine**
2. **1 participant out of 10**, **Low fine**
3. **1 participant out of 4**, **Low fine**
4. **1 participant out of 2**, **Low fine**

Treatment 16-19: Strategy method for the fine (with probability between subjects)

There is a chance that we will audit your message. We will randomly select **1 out of 20** participants and verify that the number indicated in their message corresponds to the actual outcome of the die roll (6).

If you are randomly selected and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, you will either lose all or half of your earnings.

We now ask you to choose the message you would like to send for both the possible cases. At the end of the experiment, you will be notified about what the actual fine is, and the corresponding message will be implemented.

Once you are ready, please choose the message you would like to send to the other participant for both the possible fines.

1. If you are randomly selected to be audited and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose all of the $10 participation fee and all of the payment associated to with the message you sent, leaving you with $0.25 earnings for this task.**

2. If are randomly selected to be audited and your message does not correspond to the actual outcome of the die roll, you will be fined. In particular, **you will lose half of the $10 participation fee and half of the payment associated to with the message you sent, leaving you with $12.75 earnings for this task.**

Bibliography

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

Allingham, M.G. and Sandmo, A. (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics*, 1, 323-338.

Alm, J., McClelland, G. H., and Schulze, W.D. (1992). Why do People Pay Taxes? *Journal of Public Economics*, 48, 21-38.

Amabile, T.M. (1996). *Creativity in Context: Update to the Social Psychology of Creativity*. New York: Westview Press.

Amabile, T.M, (1998). How to Kill Creativity. *Harvard Business Review*, 40, 39-58.

Amir, O., Rand, D.G., and Gal, Y.K. (2012). Economic Games on the Internet: The Effect of $1 Stakes. *PLoS ONE* , 7, e31461.

Anderson, L. R., and Stafford, S.L. (2003). Punishment in a Regulatory Setting: Experimental Evidence from the VCM. *Journal of Regulatory Economics*, 24, 91-110.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large Stakes and Big Mistakes. *The Review of Economic Studies*, 76, 451-469.

Azoulay, P., Graff Zivin, J.S., and Manso, G. (2011). Incentives and Creativity: Evidence from the Academic Life Sciences. *RAND Journal of Economics*, 42, 527-554.

Balietti, S., Goldstone, R.L., Helbing, D. (2016). Peer Review and Competition in the Art Exhibition Game. *Proceedings of the National Academy of Sciences*, 113, 8414-8419.

Barron, G., and Erev, I. (2003). Small Feedback-Based Decisions and their Limited Correspondence to Description-Based Decisions. *Journal of Behavioral Decision Making*, 16, 215-233.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: The Role of Guilt. *Journal of Economic Behavior & Organization*, 93, 227-232.

Beccaria, C. (1986). *On Crimes and Punishments*. Indianapolis: Hackett Publishing Company

Beck, P.J., Davis, J. S, and Jung, W.O. (1991). Experimental Evidence on Taxpayer Reporting under Uncertainty. *Accounting Review*, 66, 535-558.

Becker, G.S. (1957). *The Economics of Discrimination* (First Edition). Chicago: The University of Chicago Press.

Becker, G.S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, 76, 169-217.

Beckers, N., Cools, M., and Van den Abbeele, A. (2010). The Impact of Incentives on Creatvity: A Literature Review. *Review of Business and Economics*, 463-483.

Belot, M., Bhaskar, V., and Van De Ven, J. (2012). Can Observers Predict Trustworthiness? *Review of Economics and Statistics*, 94, 246-259.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience Theory of Choice Under Risk. *Quarterly Journal of Economics*, 127, 1243-1285.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and Consumer Choice. *Journal of Political Economy,* 121, 803-843.

Bott, K.M. (2016). Increasing Tax Compliance–Auditing, Appeals to Tax Morale, or Both? A Lab Experiment. *Working Paper.*

Bott, K., Cappelen, A.W., Sørensen, E.Ø., and Tungodden, B. (2014). You've Got Mail: A Randomised Field Experiment on Tax Evasion. *Working Paper.*

Bradler, C. (2015). How Creative Are You? – An Experimental Study on Self-Selection in a Competitive Incentive Scheme for Creative Performance. *ZEW Discussion Paper No. 15-021.*

Bradler, C., Neckermann, S., and Warnke, A. (forthcoming). Incentivizing Creativity: A Large-Scale Experiment with Performance Bonuses and Gifts. *Journal of Labor Economics.*

Byron, K., and Khazanchi, S. (2012). Rewards and Creative Performance: A Meta-Analytic Test of Theoretically Derived Hypothesis. *Psychological Bulletin,* 138, 809-830.

Camerer, C., and Ho, T.H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*.167-196.

Cappelen, A.W., Sorensen, E., and Tungodden, B. (2013). When Do We Lie? *Journal of Economic Behavior & Organization,* 93, 258-265.

Chalfin, A., and McCrary, J. (2013). Are U.S. Cities Underpoliced? Theory and Evidence. *NBER Working Paper No. 18815*.

Chalfin, A., and McCrary, J. (2017). Criminal Deterrence: A Review of the Literature. *Journal of Economic Literature,* 55, 5-48.

Charness, G., and Dufwenberg, M. (2006). Promises and Partnership. *Econometrica*, *74*, 1579-1601.

Charness, G., Gneezy, U., and Kuhn, M. (2012). Experimental Method: Between-Subject and Within-Subject Design. *Journal of Economic Behavior & Organization,* 81, 1-8.

Charness, G., and Grieco, D. (forthcoming). Creativity and Incentives. *Journal of the European Economic Association.*

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and Taxation: Theory and Evidence. *American Economic Review*, 99, 1145-1177.

Cropley, A. (2006). In Praise of Convergent Thinking. *Creativity Research Journal,* 18, 391–404.

Cropley, A., and Cropley, D. (2008). Resolving the paradoxes of Creativity: An Extended Phase Model. *Cambridge Journal of Economics,* 38, 355-373.

Dai, Z., Galeotti, F., and Villeval, M.C. (forthcoming). Dishonesty in the Lab Predicts Dishonesty in the Field. An Experiment in Public Transportations. *Management Science*.

DeAngelo, G., and Charness, G. (2012). Deterrence, Expected Cost, Uncertainty and Voting: Experimental Evidence. *Journal of Risk and Uncertainty*, 44, 73-100.

DePaulo, B. M., and Kashy, D.A. (1998). Everyday Lies in Close and Casual Relationships. *Journal of Personality and Social Psychology,* 74, 63-79.

Drago, F., Galbiati, R., and Vertova, P. (2009). The Deterrent Effects of Prison: Evidence from a Natural Experiment. *Journal of Political Economy,* 117, 257-280.

Dreber, A., and Johannesson, M. (2008). Gender Differences in Deception. *Economics Letters,* 99, 197-199.

Eckartz, K., Kirchkamp O., and Schunk, D. (2012). How do Incentives Affect Creativity? *Working Paper*.

Ederer, F., Manso, G. (2013). Is Pay for Performance Detrimental to Innovation*? Management Science, 59*, 1496-1513.

Ellsberg, D. (1961). Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*, 75, 643-669.

Enke, B., and Zimmermann, F. (forthcoming). Correlation Neglect in Belief Formation. *Review of Economic Studies*.

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

Erat, S., and Gneezy, U. (2016). Incentives for Creativity. *Experimental Economics*, 19, 269-280.

Evans, W., and Owens, E. (2007). COPS and Crime. *Journal of Public Economics,* 91, 181-201.

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

Fox, C.R., and Tversky, A. (1995). Ambiguity Aversion and Comparative Ignorance. *Quarterly Journal of Economics*, 110, 585-603.

Frank, B., and Schulze, G. G. (2000). Does Economics Make Citizens Corrupt? *Journal of Economic Behavior & Organization*, 43, 101-113.

Franke, N., Poetz, M.K., and Schreier, M. (2014). Integrating Problem Solvers from Analogous Markets in New Product Ideation. *Management Science*, 60, 1063-1081.

Friedland, N. (1982). A Note on Tax Evasion as a Function of the Quality of Information about the Magnitude and Credibility of Threatened Fines: Some Preliminary Research. *Journal of Applied Social Psychology*, 12, 54-59.

Friedland, N., Maital, S., and Rutenberg, A. (1978). A Simulation Study of Income Tax Evasion. *Journal of Public Economics*, 10, 107-116.

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

Gabaix, X. (2017). Behavioral Inattention. *NBER Working Paper No. 24096*.

Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review*, 95, 384-394.

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

Gneezy, U., Saccardo, S., Serra-Garcia, M., and van Veldhuizen, R. (2015). Motivated Self-Deception, Identity and Unethical Behavior. *Working Paper*.

Goodman, J.K., Cryder, C., and Cheema, A. (forthcoming). Data Collection in a Flat World: Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*.

Greiner, B. (2004). An Online Recruiting System for Economic Experiments. *Forschung und wissenschaftliches Rechnen GWDG Bericht 63*. edited by K. Kremer and Macho V. Göttingen, Gesellschaft für Wissenschaftliche Datenverarbeitung, 79–93.

Gross, D. (2016). Creativity Under Fire: The Effects of Competition on Innovation and the Creative Process. *Working Paper.*

Hagmann, D., Harman, J.L., and Gonzalez, C. (2015). Wait, Wait... Don't Tell Me: Repeated Choices with Clustered Feedback. *Working Paper.*

Harbaugh, W.T., Mocan, N., and Visser, M.S. (2013). Theft and Deterrence. *Journal of Labor Research,* 34, 389-407.

Helland, E., and Tabarrok, A. (2007). Does Three Strikes Deter? A Nonparametric Estimation. *Journal of Human Resources,* 42, 309-330.

Hertwig, R., and Erev, I. (2009). The Description–Experience Gap in Risky Choice. *Trends in Cognitive Sciences*, 13, 517-523.

Hertwig, R., Barron G., Weber, E.U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science,* 15, 534-539.

Holmstrom, B., and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset

Ownership and Job Design. *Journal of Law, Economics, and Organization*, 7, 24-52.

Holt, C.A., and Laury, S.K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92, 1644-1655.

Horton, J.J., and Chilton, L.B. (2010). The Labor Economics of Paid Crowdsourcing. *Working Paper*.

Horton, J.J., Rand, D.G., and Zeckhauser, R.J. (2011). The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics*, 14, 399-425.

Hossain, T., and List, J.A. (2012). The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Management Science*, 58, 2151-2167.

Houser, D., Vetter, S., and Winter, J. (2012). Fairness and Cheating. *European Economic Review*, 56, 1645-1655.

IRI. (n.d.). *Leading Countries by Gross Research and Development (R&D) Expenditure Worldwide in 2018 (in Billion U.S. Dollars)*. In Statista - The Statistics Portal. Retrieved September 7, 2018, from https://www.statista.com/statistics/732247/worldwide-research-and-development-gross-expenditure-top-countries/.

Jeppesen, L.B., and Lakhani, K.R. (2010). Marginality and Problem Solving Effectivness in Braodcast Search. *Organizational Science*, 21, 1016-1033.

Kachelmeier, S.J., Reichert, B.E., and Williamson, M.G. (2008). Measuring and Motivating Quantity, Creativity, or Both. *Journal of Accounting Research*, 46, 341-374.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavorial Economcis. *American Economic Review*, 93, 1449-1475.

Kahneman, D., and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263-291.

Khadjavi, M. (2015). On the interaction of deterrence and emotions. *Journal of Law, Economics, and Organization*, 31, 287-319.

Khadjavi, M. (2018). Deterrence Works for Criminals. *European Journal of Law and Economics*, 46, 165-178.

Kuziemko, I., Norton, M.I., Saez, E., and Stantcheva, S. (2015). How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments. *American Economic Journal*, 105, 1478-1508.

Laffont, J.J., and Martimort, D. (2009). *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press.

Laske, K. and Saccardo, S. (2018). Do Fines Deter Unethical Behavior? The Effect of Systematically Varying the Size and Probability of Punishment. *Working Paper*.

Laske, K. and Schröder, M. (2018). Quality through Quantity - The Effects of Piece-Rate Incentives on Creative Performance. *Working Paper*.

Lazear, E.P. (2000). Performance Pay and Productivity. *American Economic Review*, 90, 1346-1361.

Lejarraga, T., and Gonzalez, C. (2011). Effects of Feedback and Complexity on Repeated Decisions from Description. *Organizational Behavior and Human Decision Processes*, 116, 286-295.

Levitt, S.D. (1997). Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. *American Economic Review*, 87, 270-290.

Levitt, S.D. (2002). Using Electoral Cycles in Police Hiring to Estimate the Effects of Police on Crime: Reply. *American Economic Review*, 92, 1244-1250.

Lin, M.J. (2009). More Police, Less Crime: Evidence from U.S. State Data. *International Review of Law and Economics*, 29, 73-80.

Loftin, C., and McDowall, D. (1981). One With A Gun Gets You Two: Mandatory Sentencing and Firearms Violence in Detroit. *ANNALS of the American Academy of Political and Social Science*, 455, 150-167.

Mazar, N., Amir, O., and Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, 45, 633-644.

McDowall, D., Loftin, C., and Wiersema, B. (1992). A Comparative Study of the Preventive Effects of Mandatory Sentencing Laws for Gun Crimes. *Journal of Criminal Law and Criminology*, 83, 378-394.

Murphy, K. R. (1993). *Honesty in the Workplace*. Pacific Grove, CA: Brooks/Cole.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411-419.

Petters, L., and Schröder, M. (2017). Justification as a Key Determinant of the Effect of Affirmative Action. *Working Paper*.

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature*, 37, 7-63.

Rabin, M. (2000). Risk Aversion and Expected Utility Theory: A Calibration Theorem. *Econometrica*, 68, 1281-92.

Rizzolli, M., and Stanca, L. (2012). Judicial Errors and Crime Deterrence: Theory and Experimental Evidence. *Journal of Law and Economics*, 55, 311-338.

Schildberg-Hörisch, H., and Strassmair, C. (2012). An Experimental Test of the Deterrence Hypothesis. *Journal of Law, Economics, and Organization*, 28, 447-459.

Schwartzstein, J. (2014). Selective Attention and Learning. *Journal of the European Economic Association*, 12, 1423-1452.

Simon, Herbert A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69, 99-118.

Simonton, D.K. (2003). Scientific Creativity as Constrained Stochastic Behavior: The Integration of Product, Person, and Process Perspectives. *Psychological Bulletin*, 129, 475-495.

Spicer, M.W., and Thomas, J.E. (1982). Audit Probabilities and the Tax Evasion Decision: An Experimental Approach. *Journal of Economic Psychology*, 2, 241-245.

Sutter, M. (2009). Deception through Telling the Truth? Experimental Evidence from Individuals and Teams. *Economic Journal*, 119, 47-60.

The Conference Board (2012). *CEO Challenge 2012*, Research Report R-1491-12-RR.

The Conference Board (2013). *CEO Challenge 2013*, Research Report R-1511-13-ES.

The Conference Board (2014). *CEO Challenge 2014*, Research Report R-1537-14-RR.

Tversky, A., and Kahneman D. (1973). Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, 5, 207-232.

Tversky, A., and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5, 297-323.

Van de Ven, J., and Villeval, M.C. (2015). Dishonesty under Scrutiny. *Journal of the Economic Science Association*, 1, 86-99.

Webley, P. (1987). Audit Probabilities and Tax Evasion in a Business Simulation. *Economics Letters*, 25, 267-270.

Wiley, J. (1998). Expertise as Mental Set: The Effects of Domain Knowledge in Creative Problem Solving. Memory & Cognition, 26, 716-730.

Wu, G., and Gonzalez, R. (1996). Curvature of the Probability Weighting Function. *Management Science*, 42, 1676-1690.