Lohse, Tim; Konrad, Kai A.; Qari, Salmai

Conference Paper

Deception Choice and Audit Design - The Importance of Being Earnest

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Taxation III, No. C15-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Lohse, Tim; Konrad, Kai A.; Qari, Salmai (2014) : Deception Choice and Audit Design - The Importance of Being Earnest, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Taxation III, No. C15-V3, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/100577

Deception Choice and Audit Design — The Importance of Being Earnest[☆]

Kai A. Konrad[a,b], Tim Lohse[a,c,*], Salmai Qari[a]

[a] *Max Planck Institute for Tax Law and Public Finance, Marstallplatz 1, 80539 Munich, Germany*
[b] *Social Science Research Center Berlin, Reichpietschufer 50, Berlin*
[c] *Berlin School of Economics and Law, Alt-Friedrichsfelde 60, 10315 Berlin, Germany*

Abstract

We study deception choices and deception detection in a tax compliance experiment. We find large systematic differences in individual deception abilities. Tax payers are conscious about their own deception abilities. The empirical outcomes are in line with a theory suggesting that tax payers make their choices whether to underreport or report truthfully on the basis of their own deception ability. Tax payers with high deception ability are more likely to underreport. This selection effect is stronger if the fines for underreporting are higher. These results provide an (additional) reason why random audits are superior to audits based on discretionary choice.

JEL Classification Codes: H31, K42, C91

Keywords: lie-catching, self-selection, tax compliance, deception

1. Introduction

Individuals have a choice whether to lie or to tell the truth. This choice depends on a number of issues. One important aspect is how individuals assess their own subjective probability that their deception would be detected.

[☆] A preliminary and short version of this paper with a brief overview about some results has been circulated under the title 'Deception Detection and the Role of Self-Selection'. This paper has a new title, but replaces this previous version.

[*] Corresponding author

Email addresses: kai.konrad@tax.mpg.de (Kai A. Konrad), tim.lohse@hwr-berlin.de (Tim Lohse), salmai.qari@tax.mpg.de (Salmai Qari)

Detection may occur through a specific audit mechanism, such as a face-to-face control by an inspector. We conduct a tax compliance experiment combined with a lie-catching experiment to study the role of this probability assessment for deception choices. We study, first, whether individuals differ in their deceptive abilities and try to measure these differences. Second, we analyze if the individuals' self-assessed audit probabilities are correlated with the inspectors' assessments of them being honest or not. Third, if individuals base their deception choices on their self-assessed deceptive abilities, we ask if there is evidence for self-selection by which more capable liars are more likely to lie. Thereby, we also consider how self-selection depends on the incentive structure, and on the fines for detected deception. From these insights gained, we draw conclusions about the optimal design of the audit mechanism.

Choice problems for which the attempt to deceive involves a material risk are rather common. When filing for income taxes, people may report all their income truthfully. Alternatively, they may underreport their income. Underreporting leads to lower tax payments if it remains undetected, but triggers a fine and leads to higher payments if it is detected.[1] Inside an organization people may be upfront about what went wrong and may apologize, or they may apply deceptive strategies and try to avoid taking responsibility for mistakes.[2] Job applicants may report truthfully or lie about their competencies and skills.[3] Salesmen may report truthfully, or inflate their expense claims or shirk on their working hours.[4] And supervisors may misreport the perfor-

[1]This choice between a safe tax payment outcome and a gamble involving a possible fine is at the heart of much of the tax compliance literature (see, e.g., Allingham and Sandmo 1972 or more advanced models such as Rheinganum and Wilde 1985).

[2]See, e.g., Kellerman (2006) for a discussion of the high stakes for corporate leaders and the optimal choice problem of whether to apologize or to deny or remain silent.

[3]Deception in employment interviews has attracted considerable interest among social psychologists. A survey and meta-study is by Barrick, Shaffer and DeGrassi (2009).

[4]Such choice problems are the starting point of the literature on efficiency wages (Shapiro and Stiglitz 1984) or the principal-agent literature on moral hazard.

mance of their workers.[5] In each of these cases individuals choose whether to report truthfully, or to attempt to deceive their counterparts.

To study individual deception ability and its role for deception choices we conduct a lie-catching experiment that is framed in the context of tax compliance with 231 subjects - "judges" - who rate videotaped tax declarations of 80 subjects - "tax payers" leading to 9240 observations. These videotapes were taken from a tax compliance experiment. The tax payers had a face-to-face interview with an interviewer who had the role of a tax inspector in a laboratory environment. Tax payers had to decide whether to underreport taxes, or to pay taxes truthfully. They knew about the possible monetary upsides and downsides of underreporting compared to truthful compliance: underreporting was rewarding if they were not caught, and more costly than truthful compliance if they were caught. The data show tax payers in two different treatments. One treatment had high fines and a second one had low fines. The treatment difference in fines allows us to focus on self-selection among the tax payers. A high fine should discourage underreporting in general, and it should discourage low-ability liars more effectively than high-ability liars. For this reason, the sets of individuals who choose to underreport income in the two treatments should have different deception abilities. Selection, and, hence, average deception ability should be higher among individuals who choose to underreport in the treatment in which underreporting is discouraged by higher fines.

The answers to the research questions outlined above are: First, tax payers exhibit systematic differences regarding their probability of being correctly classified. This heterogeneity is found for both truthful and underreporting tax payers. There are underreporting tax payers who are consistently classified as dishonest (measured by a high number of judges who classify them as untruthful) and there are underreporting tax payers who are systematically

[5]See, for instance, the experimental evidence by Rosaz and Villeval (2012).

wrongfully classified as truthful by a majority of judges. And similar heterogeneity exists for honestly reporting tax payers.[6] Second, subjects can to some extent correctly assess how truthful they are perceived by others. We find subjects' self-assessed likelihood for an audit to be positively correlated with their dishonesty scores as stated by the interviewers. The dishonesty score for each subject is equal to the fraction of interviewers who assess this particular subject as dishonest, i.e., the score is equal to zero if all interviewers assess this subject as honest and equal to one if all interviewers assess this subject as dishonest. Thus, the self-assessed deception ability influences tax compliance choices. We find several pieces of evidence confirming this. A first indication is the low lie-catching rates: the underreporters manage to be detected with a probability lower than the probability that could be obtained by a pure random audit device. They manage to look more truthful on average than the truthful low-income earners with whom they are pooled. Further, fewer tax payers underreport if the fines for detected deception are higher, and these fewer underreporters have an even higher deception success rate than underreporters in a treatment with low fines. As we will argue, this is evidence for self-selection based on self-assessed deception ability.

Our findings have implications for the design of auditing procedures which, generally speaking, can be based either on face-to-face contact with discretion about who to audit, or on a pure random mechanism. The choice of procedure may cause different psychological costs for the individuals who may be audited, including feelings of anxiety, ambiguity or a general uneasiness. Also, the size of transaction costs may differ, and, as a major

[6]Perhaps surprisingly, the judges do not show systematic differences in their ability to detect liars. Ekman and O'Sullivan (1991) found differences in accuracy for deception detection among occupational groups. However, the issue whether experienced lie-catchers have higher detection rates remains controversial among psychologists. In particular, experience seemingly loses much impact if the assessment context is changed. We do not contribute much to this controversy, as all our subjects are students. We find that the heterogeneity in lie-catching ability among the student judges is low.

disadvantage of face-to-face contact, an encounter between an auditor and the individuals who report may facilitate undesirable collusion between the parties. As analyzed theoretically by Chander and Wilde (1992), Hindriks, Keen and Muthoo (1999) and Ksh (2008), corrupt tax inspectors may accept bribes or extort money from tax payers. Direct personal contact and unrecorded communication may simplify or may even be a prerequisite for bribery. On the other hand, an advantage of face-to-face contact and direct communication is the potential of more successful detection of misreporting. This would be an important justification for such costly audit procedures. However, our results do not provide support for the benefits of personal contact as they suggest that personal contact and subjective assessments that are based on face-to-face communication may be inferior to strict random audits – even disregarding their direct and indirect cost of implementation – for two reasons. First, the lie-catching ability of assessors is seemingly low. Second, we draw attention to the self-selection forces among tax payers as a response to auditing procedures. High-powered incentives seemingly cause stronger selection, such that deception is used only by individuals who have high deception skills. An implication is that a non-randomized audit mechanism, where people have discretion about who to audit, may perform worse than a random audit. Discretionary decision making discourages deception by weak liars more effectively, and it may discourage liars of superior deception ability less effectively. As a result, the set of individuals who lie consists of individuals who have superior deception abilities. This result also speaks to Becker's (1968) theorem on the optimality of maximum fines. High fines may invite subjects with superior deceptive abilities. The composition of subjects from which an auditor has to choose may be adversely chanced.

2. Related Literature

Considerations about deception and deception detection date back to Darwin (1872), Lombroso (1876) and Freud (1959). A milestone in the experi-

mental work on lie-detection was conducted by Ekman and Friesen (1974). Since then, research on lie-catching and deception detection has been analyzed in more than 200 experiments, mostly by social psychologists. Much is known by now about deception and the ability to detect lies. Overall, much evidence suggests that the ability to detect lies is limited, but controversy about this continues. For recent surveys and meta-studies see DePaulo et al. (2003), Bond and DePaulo (2006), Vrij (2008) and Hartwig and Bond (2011). Much of this literature has concentrated on what are the cues that subjects use to detect liars, whether different status groups, and interrogation experts or professionals in particular, have a higher ability to detect lies. Whether or not to use a deception strategy typically was not a matter of choice in the experiments; people were often told to deceive. This is one of the points of criticism by DePaulo et al. (2003, p. 106) in their meta-study. Other important points are the lack of incentives and of major types of feedback. Our set-up takes into account these points. First, the subjects ("tax payers" as well as "judges") earn money if they are successful. Second, all persons which are seen in the videotapes perform an action which they have chosen, based on their monetary incentives, their true taxable incomes and their perceptions about their own deception abilities. This is an important departure, as it may potentially lead to self-selection: the more capable liars may choose to deceive. As a consequence, the "quality of liars" is different as compared to a setting in which all subjects who lie are forced to, or advised to lie, and this self-selection is presumably causal for the low lie-detection rate. In addition, we study the effect of changes in the monetary incentives for the self-selection among subjects according to their deceptive abilities. Self-selection can also explain our finding that the rate of correct judgements was even lower for tax payers if the punishment for underreporting that was detected was higher.

Our analysis is also related to several lines of literature in economics. The management science literature addressed a number of aspects of audit

design. Yim (2009), for instance, discusses audit sampling plans and their relationship with the audit budget. Erat and Gneezy (2012) highlight the role of consequences for others for individual choice of a deception strategy. Brandts and Charness (2003) conducted an experiment on the role of deliberate deception for the willingness to exercise costly punishment. Fagart and Sinclair-Desgagné (2007) and Dionne, Giuliano and Picard (2009) study the design of monitoring systems in dynamic contexts. From a perspective of theory, Crawford (2003) considers the strategic incentives for deception. Holm (2010) considers signalling and signal extraction if the recipient of the signal has (and is believed to have) a probabilistic truth telling detection technology.

The importance of an individual's face in a situation of economic interaction has been pointed out by Eckel and Petrie (2011). They conduct a trust game experiment in which individuals are allowed to buy a photo of their counterpart beforehand. From such a photo individuals may infer some characteristics that have been identified to play an important role such as beauty (Mobius and Rosenblat (2006), Wilson and Eckel (2006)), ethnicity (Habyarimana, Humphreys, Posner and Weinstein (2007)), gender (Solnick and Schweitzer (1999), Andreoni and Petrie (2008)) or race (Castillo and Petrie (2010)). Similarly, Eckel and Petrie (2011) find the informational value of a face to be non-zero and observe a change in economic behaviour once the veil of anonymity is lifted. Konrad, Lohse and Qari (2013) consider the role of face value in a tax compliance game.

Experimental work on whether individuals can unveil incomplete information in a strategic situation that involves face-to-face communication was carried out by Frank, Gilovich and Regan (1993), Brosig (2002), Ockenfels and Selten (2000), Sánchez-Pagés and Vorsatz (2007) and Holm and Kawagoe (2010). The first three papers consider strategic interaction with face-to-face contact. They ask whether the veil of incomplete information about each other can partially be lifted by the fact that individuals see each other face-

to-face, and see their actions. Ockenfels and Selten (2000) study a bargaining context with face-to-face interaction and incomplete information. They find that subjects' bargaining offers in the course of bargaining provide cues about players' types.[7] Holm and Kawagoe (2010) consider an experiment which resembles a matching-pennies game and in which players earn money if they can correctly assess whether their counterpart lies or tells the truth. In their set-up, this counterpart has an incentive to choose a mixed strategy and mix lying and truth-telling equally in the theory equilibrium. None of these experiments focus on the role of heterogeneity in deception ability for the choice about deception and for the self-selection of players by which only the players with higher deceptive abilities use deception.

3. Methodology

We use video clips of tax compliance interviews which were generated in the context of a tax compliance experiment. A randomly selected subset of these videotapes was shown to a large number of students whose task was to assess which videotape shows a liar and which shows a truth-teller. These lie-catching interviews are the core of the experiment which we report about in this paper.[8] However, it is important to get a clear picture about the set of videos we used. Therefore, we first explain the experimental conditions and the process that led to the compliance videotapes in greater detail. Then we describe the design of the actual lie-catching experiment that draws on these videos. We discuss the two-level structure of the data with the potential sources of heterogeneity. Further, we discuss the role of heterogeneity for self-selection, and how heterogeneity interacts with a change in the incentives to

[7] Even though Ockenfels and Selten (2000) is not a lie-catching experiment, it relates to our study. A major difference with what we do is that the hidden characteristics of individuals led them to different, seemingly informative behavior (immediate consensus versus bargaining delay - in their set-up).

[8] The use of videoclips for lie-catching studies is common and traces back to Ekman and Friesen (1974).

lie.

3.1. The compliance interview clips

The video clips have been produced in the context of an economic experiment on tax compliance that was conducted at MELESSA, the experimental laboratory at the University of Munich in March 2012. Each video shows the tax declaration of a "tax payer" as part of a short standardized dialogue face-to-face to a person with the role of a "tax inspector". Each video took about 20 seconds and the questions and answers followed a strict protocol. The tax payers were students recruited by the MELESSA laboratory in Munich using the software ORSEE (Greiner 2004); tax inspectors were student assistants of the Max Planck Institute.[9]

Subjects' true laboratory income was assigned to them. This income was either high (1000 Taler) or low (400 Taler). In the compliance dialogue, a tax payer could claim to have nothing to declare (meaning that he or she has low income resulting in a zero tax liability), or declare high income. No taxes had to be paid on low income, whereas declaring high income triggered a positive tax liability (200 Taler). Tax payers with low income had a unique best choice: declare that they have low income. Empirically they behaved in line with this dominant strategy. Tax payers with high income had to make the choice whether to report truthfully and pay a tax, or to underreport. If they reported truthfully they paid 200 Taler in taxes. If they reported low income, half of them received an audit. The outcome of who received an audit was influenced by student research assistants who performed the task as tax inspectors.[10] If an audit took place, the audit revealed with certainty

[9]In a situation with face-to-face communication deceptive behavior may invoke specific psychological effects such as higher mental cost of lying (Vanberg 2008, Lundquist, Ellingsen, Gribbe and Johannesson 2009), shame (Coricelli, Joffily, Montmarquette and Villeval 2010) or guilt aversion (Charness and Dufwenberg 2006, among others). Konrad, Lohse and Qari (2012) also confirm that face-to-face interaction has a (weak) effect for deception choices compared to a fully anonymous computerized treatment.

[10]More specifically, the tax inspectors had to assess a series of declarations and had

whether or not the person underreported. A person who underreported then had to pay the tax plus a fine. This fine was small (100 Taler) in some of the sessions. We refer to these sessions as the "low-fine treatment" sessions. The fine was high (300 Taler) in other sessions, and we refer to these as the "high-fine treatment" sessions.[11]

3.2. The assessment experiment

For the assessment interviews we used a randomly composed subset of the compliance video clips that show persons who truthfully report low income and of clips of persons who have high income but underreport and claim to have low income. This generates samples of clips showing subjects who all make the same statement: they claim that they have low income. Some tell the truth while others lie.[12] We used 80 clips, each clip showing a different tax payer. Of these, 40 clips show low-income tax payers who truthfully declared low income and paid no taxes. Furthermore, 40 clips show individuals who had high income but made the choice to underreport. Of these, 20 clips were randomly picked from the underreporters who underreported in the low-fine treatment, and 20 videos showed underreporters in the high-fine treatment.

We partitioned the 80 video clips in two disjoint subsets of 40 videos, consisting of 20, 10 and 10 videos, respectively. These were shown to students

to rank the tax payer subjects who declared low income with respect to whether they perceived them as more or less honest. This ranking affected who of the underreporting tax payers received an audit. Accordingly, tax payers made their declaration choice knowing that their appearance affected their audit probability.

[11]Besides the fine and the audit probability, other determinants have been identified to affect individuals' compliance decision. These include intrinsic motivation (Frey 1997), an inclination for pro-social behavior (Frey and Torgler 2007), fairness considerations (Hartner, Rechberger, Kirchler and Schabmann 2008), religiosity (Torgler 2006), and patriotism (Konrad and Qari 2012) among others. Andreoni, Erard and Feinstein (1998) and Slemrod (2007) provide in-depth surveys of this large literature.

[12]There were also compliance interviews in which subjects with high income declared high income. These were not useful and not used for the assessment experiment: these clips show people who always truthfully report high income. They were trivially distinguishable from clips showing individuals who (truthfully or falsely) report low income.

whose task was to assess the truthfulness of persons shown in these clips. We refer to these students as "judges". In total, 231 students were invited to the laboratory at the Technical University of Berlin for this purpose in November 2012.[13] Judges were from diverse fields of study. One set of 40 videos was shown to 120 judges, the other set was shown to 111 other judges. These students were grouped in 10 sessions of up to 24 participants each, reflecting the capacity in the laboratory.

The judges were told that they will see a sequence of 40 clips with tax compliance dialogues on their computer screens, and roughly how these videos were produced and what they show. Judges did not receive any additional information about which video came from which treatment. In fact, they did not even receive information about the fact that the video clips emerged from two different treatments, one with low fines and one with high fines. However, we informed judges that the share of truthful reports among the 40 videos was about one half. Each judge watched the forty clips on the computer screen and had a headphone to listen to the tax payers' reports. The videos were shown in a random order. Judges were not allowed to return to previous clips they had already assessed and change their judgement in the course of the experiment.

Judges had monetary incentives to make correct judgements. Out of the 40 assessments of a judge, the computer randomly selected five rounds for payment. This was in order to provide them with a stronger feeling that their judgement matters and to make a simple hedging strategy less attractive by which subjects may simply rate the first half, or every uneven video, as truthful. Judges were paid EUR 5 for each correct assessment among these five assessments that were selected to be paid for, and they received a show-

[13]These students were in the subject pool of the TU lab in Berlin. Subjects in the compliance videos were students at the University of Munich. This makes an overlap of subjects almost impossible. All participants were recruited using the software ORSEE (Greiner, 2004).

up fee of EUR 5. Accordingly, realized final payments were between EUR 5 and EUR 30 with an average of EUR 17.99 (SD=6.23).

3.3. Theory predictions

Our set-up allows to inquire into the heterogeneity of tax payers and the implications of this heterogeneity for their behaviour. We ask: do tax payers differ in their deception ability? And do they know about their own ability? We further ask: if tax payers differ in their deception ability, how should this heterogeneity affect their choice behaviour? How is their decision whether to underreport affected by their own ability? How is this relationship between their own deceptive ability and choice affected by different monetary disincentives for underreporting?

From a decision theory point of view, consider the specific choice problem. Let there be two possible levels of income: Y_0 and Y_1, with $Y_0 < Y_1$. Let the statutory taxes be $T(Y_1)$ and $T(Y_0)$ with $T(Y_1) > T(Y_0) = 0$. Consider a tax payer who has high income Y_1. The tax payer may declare this high income truthfully and pay a tax $T(Y_1)$. In this case the final income is $Y - T(Y_1)$. If the tax payer underreports, given that there are only two possible levels of income, the tax payer declares low income Y_0. The statutory tax on low income is $T(Y_0) = 0$. The tax payer knows by design and in the aggregate, a share p of the persons who underreport receive an audit. And in the experiment that led to the video clips, tax payers were explicitly informed about this aggregate audit rate. If the tax payer falsely reported low income, and receives an audit, the final income is $Y_1 - T(Y_1) - D$, where D is a monetary fine that can be either high or low. If the audit probability is p and exogenous, then, a tax payer who maximizes his monetary payoff reports truthfully if $(T(Y_1) + D)p > T(Y_1)$ and underreports if $(T(Y_1) + D)p < T(Y_1)$. Even if p is objectively and exogenously given, other considerations such as risk attitudes, or other behavioural attitudes or a mental benefit or cost from truth telling or lying may make individuals deviate from this

decision rule.[14] Statistically speaking, however, we would expect that fewer individuals choose to underreport if the fine D is higher.

The individual tax payer may assess his or her own probability for an audit and conclude that their self-assessed subjective audit probabilities deviate from the average audit probability. If this is the case, a tax payer's self-assessed ability, but also the expectations about the deception abilities of other tax payers and their choice behaviour matter for each single tax payer's assessment of their own subjective audit probability. The outcome could be characterized as an equilibrium of a Bayesian game that can be established if each tax payer knows their own deception ability and the distribution of deception abilities from which other tax payers' abilities are drawn. We do not outline this game in full. But a tax payer's own audit probability should be a function of self-assessed deception ability in equilibrium in this case, and it should hold that a tax payer's own subjective audit probability is decreasing in self-assessed deception ability: higher own ability does not change the benefits of truth telling, but increases the benefits from underreporting. Accordingly, tax payers who have a higher self-assessed deception ability should be more inclined to underreport.

Recall that we have two treatments of the tax compliance game. One has a low fine, the other has a high fine. For the low fine treatment, the decision to underreport may pay off in expected value terms, even if the subjective audit probability is slightly higher than 1/2. Individuals who think that their deceptive ability is very low may still prefer truth telling, but individuals with a medium or high deception ability may prefer to underreport. For the high fine treatment, the decision to underreport pays off in expectation only for a sufficiently low own subjective audit probability. Individuals who

[14]The tax payers' decisions may also depend on other, also unobserved aspects. These other dimensions have several possible underpinnings. Sánchez-Pagés and Vorsatz (2007) consider possible norms about truth-telling; according to Gneezy (2005) a subject may prefer truth-telling, but may lie if other reasons make lying attractive.

think that their deception ability is just average or below average may prefer truth telling. Only individuals who think that their deception ability is sufficiently high may prefer to underreport in the high-fine treatment. This consideration leads to suggestions about the number of underreporters and about the composition of underreporters in our experiment. We expect that fewer individuals choose to underreport in the high-fine treatment. And we expect that, on average, the members of the group of underreporters in the high-fine treatment have higher deception ability than the members of the group of underreporters in the low-fine treatment.

To summarize these considerations, we formulate three hypotheses:

- *Hypothesis 1*: We ask if there are systematic differences between tax payers with respect to how they are assessed by the judges. Are there tax payers who are judged as honest systematically more often than average and others who are judged as being dishonest systematically more often than average? For a theory that bases the choice of whether or not to choose a deception strategy on differences about deceptive ability, the existence of systematic heterogeneity as regards this dishonesty score is an important pre-requisite and one of the most fundamental building blocks.

- *Hypothesis 2*: Suppose such differences in dishonesty scores exist. Deception choices are made by the tax payers and not by the judges. Therefore it is important that tax payers themselves are aware of these differences, or at least have other means to base their choices on their deception abilities. To explore this we consider if tax payers' own assessments about their subjective audit probabilities are correlated with the judges' assessments. As a measure of their self-assessment at the end of the compliance experiment, tax payers were asked whether they think their probability for receiving an audit was below 50 percent, above 50 percent or equal to 50 percent. While for several reasons this

is not the perfect variable to measure self-assessed deception ability, our theory suggests that this measure is positively correlated with the dishonesty score of tax payers.

- *Hypothesis 3*: If tax payers differ in their dishonesty score and are aware of these differences, we can ask what is the relationship between a tax payer's dishonesty score and the compliance decision. We expect that tax payers with higher deception ability are more inclined to underreport. We also expect that this self-selection effect is stronger in an environment in which deception detection leads to a higher fine. For an empirical assessment, we can use the treatment differences in fines to test this prediction. We expect that the tax payers who underreport in the high-fine treatment have lower dishonesty scores than tax payers who underreport in the low-fine treatment.

4. Results

As described above, one set of 40 tax payer videos was assessed by 120 judges, while the second set of 40 videos was assessed by 111 judges. This led to 9240 judgements in total. The heterogeneity on the tax payer level and the judge level generate the two sources of variation that we exploit to examine our main research questions.

The hit rate, i.e., overall number of correct judgements compared to the total number of judgements, was 47.35 percent. The correct share of judgements on truthful reports was 48.23 percent, on judgements on clips that showed a person underreporting was 47.66 percent in the treatment with low fines and 45.28 percent for underreporters in the high-fine treatment. These numbers fall below a hit rate of 50 percent that emerged from a simple pure random device. This deviation may be surprising because judges could have used a simple randomization mechanism that would have led to improved hit rates. The overall hit rate of 47.35 percent is close to, but outside the

boundary of previous findings surveyed in the meta-study by Bond and DePaulo (2006). Figure 1 shows a modified plot from Bond and DePaulo (2006, p. 222) and depicts the relationship between sample size and the measured hit rates; the center of the large red square represents the coordinate with 9240 observations and a hit rate of 47.35 percent from our experiment.

Figure 1 about here

In the following, we sequentially test the three hypotheses outlined in the theory section. We first analyze whether the tax payers have systematically different dishonesty scores, and we quantify the extent of this heterogeneity. We then consider how this heterogeneity squares with self-assessed audit probabilities. Then we turn to the evidence regarding the self-selection hypothesis 3.

4.1. Tax payer heterogeneity

A first way to assess the heterogeneity of tax payers' deception ability as stated in hypothesis 1 is to consider the hit rate, i.e., to count for each tax payer how often she/he was correctly assessed. Figure 2a and 2b depict the frequency distribution that emerged from the experiment for sample 1 and 2, respectively. E.g., some tax payers were assessed correctly only in 30-40 of the more than 100 assessments, and very few tax payers were correctly assessed in about 90-100 of the 111 or 120 assessments, respectively.

Figure 2a and 2b about here

At first glance, these distributions are compatible with two very distinct processes that generate these outcomes: In the first process there would be little or no systematic heterogeneity between tax payers as regards the probability that a tax payer is assessed correctly, such that this hit probability is constant for all tax payers and close to the overall hit rate of sample 1 and

2 of 47.35 percent. In this scenario, the variation in the number of hits would reflect a sampling error. For instance, if judges simply randomized in their assessments, or if there is considerable noise in their assessments, this would generate a frequency distribution that follows a binomial distribution. In the second process, the observed heterogeneity in the hit rates across tax payers reflects systematic differences in the tax payers' probability of being assessed as being honest.

We obtain a first intuitive indication in favor of the second process to be relevant by a comparison of the frequency distributions with the probability distributions that emerged from pure random choice. As noted above, if each tax payer is correctly assessed with the same probability, then the number of hits follows a binomial distribution. The binomial distributions have a constant probability equal to the average hit rate of 47.35 percent and an associated parameter of 120 and 111, respectively, and are displayed in figure 2a and 2b. As a comparison of each binomial distribution with the observed frequency distribution in both figures reveals, the probability that the latter is a realization of the former is small for both samples. In sample 1, for instance, the theoretical probability to observe only up to 40 hits in a sample of size 111 is roughly equal to one percent. However, Figure 2a shows that the fraction of tax payers who are assessed correctly by up to 40 of the 111 judges equals almost twenty percent. Thus, the observed variation in the number of hits is simply too large to be compatible with a uniform hit probability for all tax payers.

A second descriptive technique to discriminate between the two above-mentioned scenarios is as follows. Consider a single video clip of an under-reporting person that has been shown to 120 judges, of which, say, $m > 60$ judges said that the subject's report is likely to be truthful. This individual has a below average dishonesty score. Is this score and the deviation from 60 simply an outcome of noise, or is this a systematic effect? To analyze this, one can separate the total number of assessments for this person into

the assessments of the first half of the judges and into the assessments by the second half of judges. Suppose m_1 judges among the first 60 judges declared that the individual looks trustworthy, and $60 - m_1$ judges declared that the individual looks dishonest. Let these numbers be m_2 and $60 - m_2$ for the second half of the judges. Evidently, $m_1 + m_2 = m$. If the assessment of the individual is simply noise, then m_1 and m_2 should be uncorrelated. However, if the effect that causes the positive rating of the individual is systematic, then m_1 and m_2 should be positively correlated.

Figure 3 about here

Figure 3 shows a scatter-plot that emerges if we plot the corresponding hit rates for the pairs $\frac{m_1}{n}$ and $\frac{m_2}{n}$ for all 80 subjects into the same diagram with $\frac{m_1}{n}$ and $\frac{m_2}{n}$ on the two axes, where n is the total number of judgements for this individual, that is, $n = 120$ for half of the tax payers and $n = 111$ for the other half. The plot uses a decomposition into the first half and the second half of judgements. Other decompositions could also be used. The positive correlation that emerges in Figure 3 provides a second descriptive indication for systematic heterogeneity of tax payers. It suggests that judges to some extent agree regarding their assessment of the different tax payers. In turn, this agreement implies that tax payers differ systematically regarding their dishonesty scores.

We now employ mixed-effects models that allow to quantify the extent of heterogeneity on the tax payer level to analyze hypothesis 1 econometrically.[15] The models also incorporate the possibility of systematic heterogeneity of judges in their assessment abilities. Let y_{ij} denote realizations of the 9240 assessments where y_{ij} is equal to one if tax payer j is classified

[15]The term "mixed effects model" refers to the fact that both fixed effects, e.g. dummy variables or demographic variables, and random effects for the unobserved heterogeneity are estimated.

correctly (both for liars and truthful reporters) by judge i. Let $\mathbf{x_j'}$ denote a vector of explanatory variables. These variables include in particular dummy variables that indicate in which treatment tax payer j participated, but also demographics, e.g. the tax payer's gender. $\boldsymbol{\beta}$ is the vector of fixed effects. In addition, there are two random effects: u_i is a judge-specific random intercept and v_j is a tax payer-specific random intercept. The first set of models are logistic mixed-effects models that predict the hit probability for a tax payer-video as follows:

$$prob(Y_{ij} = 1 \,|u_i, v_j) = f(\mathbf{x}_j'\boldsymbol{\beta} + u_i + v_j) \tag{1}$$

where $f(\cdot)$ is the logistic cumulative distribution function.

Intuitively, this logistic regression provides estimates for an unobserved linear-additive score that describes how easily tax payers are correctly classified. For instance, assume that older tax payers are easier to read than young ones. In this case, the coefficient for a variable *Age* would be positive and the size of the *Age*-coefficient describes the linear relationship between age and the unobserved score. The score can be negative or positive and the logistic distribution function transforms this score to a probability. Since the logistic distribution is symmetric around zero, a tax payer with a score of zero has a predicted hit probability of 50 percent.

For ease of exposition, we now rewrite the regression equation by explicitly referring to the treatment variables. In this formulation, the vector of explanatory variables $\mathbf{x_j'}$ contains all remaining variables:

$$\begin{aligned} prob(Y_{ij} &= 1 \,|u_i, v_j) \\ &= f(b_0 + b_1 HighPenalty + b_2 Liar \\ &\quad + b_3(HighPenalty \times Liar) + \mathbf{x}_j'\beta + u_i + v_j) \end{aligned}$$

The hit probability is a function of the 2×2 possible treatment conditions emerging from the two possible conditions in each of two dimensions: penalty

size, and truthfulness. Recall that only subjects are included in the sample who report a low income/endowment, such that, for a subject from this sample, filing a dishonest report is equivalent to having a high endowment and filing a truthful report is equivalent with having low income/endowment. The dummy variables modeling the treatment conditions are coded as follows: *High Penalty* is equal to one if the tax payer video clip is from the high penalty treatment. *Liar* is equal to one if the tax payer's true endowment is high and equal to zero if the true endowment is low. Finally, there is an interaction term (*High Penalty* $\times$ *Liar*). This term is equal to 1 if the video clip is showing a tax payer in the high-penalty treatment condition whose true endowment is high. This coding scheme implies that the omitted reference group is composed of those videos showing truthful reports (i.e., low endowment subjects) in the low penalty treatment condition. We first present the main results from a regression without further control variables, but we will discuss these controls jointly later.

While the fixed effects model reports the average score for the respective treatment condition, the normally distributed tax payer-specific random intercept v_j allows tax payer j to deviate from this average score. Both negative and positive values are possible and a positive intercept implies that the specific hit probability of tax payer j is larger compared to an average tax payer. The random intercepts follow a normal distribution with mean zero such that the standard deviation is the only parameter left to estimate. Therefore, the coefficients of main interest are the two parameters modeling the heterogeneity on the judge level and the individual tax payer level: σ_u is the estimated standard deviation for the judge-specific random intercepts, while σ_v is the corresponding estimate for the tax payer intercepts.

Table 1 about here

The first column of Table 1 compiles the results. The estimate for σ_v is roughly equal to 0.28. Recall that this parameter implies that the tax

payer-specific random intercepts v_j are normally distributed with mean zero and the estimated standard deviation of $\sigma_v = 0.28$. Applying the usual "2-σ-rule", 95% of the tax payer-specific random intercepts lie within the interval $[-0.56, +0.56]$. In turn, this large interval translates into a wide range of tax payer-specific hit probabilities. For example, the underreporting tax payer from the low penalty treatment condition with the smallest intercept from this interval is detected with probability $f(\hat{b}_0 + \hat{b}_2 - 0.56) = 0.34$.[16] The underreporter having the largest intercept is detected with probability $f(\hat{b}_0 + \hat{b}_2 + 0.56) = 0.61$. Similar calculations apply for the other treatment conditions.[17] To summarize, there is a considerable amount of tax payer heterogeneity. The tax payer-specific hit probabilities cover a range of about 27 percentage points between the maximum and the minimum. These findings quantitatively confirm the impression from the descriptives that the observed variation in the number of hits is too large to be compatible with a uniform hit probability for each tax payer.

It is useful to consider also the possibility of systematic heterogeneity on the judge-level. However, the corresponding estimate for the judge-specific random component (σ_u) is equal to zero. Thus, the estimation indicates that the subject pool of judges shows no systematic variation in their deception detection abilities. The heterogeneity in lie-catching ability among the student judges is low and they perform poorly.[18]

[16] Recall that for the logistic mixed effects model, $f(\cdot)$ is the logistic cumulative distribution function (cdf). Thus, the predictions are obtained by simply evaluating the logistic cdf.

[17] Honest tax payers from the low penalty condition with the smallest/largest intercept are correctly detected with probability $f(\hat{b}_0 \pm 0.56)$. Once again, this corresponds to a range of roughly 27 percentage points between the maximum and minimum hit probability of 57.25% and 30.40%, respectively. The interval for underreporting tax payers' hit probabilities from the high penalty condition is $[31.83\%, 58.86\%]$. The interval for honest tax payers' hit probabilities from the high penalty condition is $[38.94\%, 66.15\%]$.

[18] As mentioned earlier, the issue whether experienced lie-catchers have higher hit rates is essentially a research topic for psychologists and remains controversial among them. Ekman and O'Sullivan (1991) found differences in hit rates for deception detection among oc-

One concern is that the relationships are the result of unobserved variables that influence both explanatory and explained variables. Of course, we cannot rule this out completely. Column (2) of Table 1 checks whether the heterogeneity of tax payers can be explained by observable characteristics, including, for example gender or age. Using all socioeconomic characteristics we have, however, the small difference in the loglikelihoods indicates that these further characteristics practically have no explanatory power.[19] The estimated standard deviation of the tax payer-random intercept and the estimated treatment coefficients are also similar across the two models. This indicates that both the small difference in the average detection probabilities across the penalty conditions and the estimated tax payer heterogeneity are unrelated to these additional control variables.

In summary, the analysis shows that it is important to account for the two-level structure of the data. By simply inspecting the number of hits for each tax payer (Figures 2a and 2b), it is not obvious whether the variation shown in the figures reflects noise or heterogeneity on the judge-level and/or on the tax payer level. The mixed-effects model takes the data structure into account and shows that the variance component associated with tax payers is fairly large. The judges' assessments of the underreporters are highly consistent across judges. The descriptive and quantitative results are in line with the hypothesis that subjects differ in their deception abilities: some individuals are - in a probabilistic sense - perceived by others as being honest, other individuals are perceived as dishonest, and this heterogeneity

cupational groups ranging from 53% (university students) to 64% (Secret Service agents). Subsequent literature shows that experience seemingly loses much impact if the assessment context is changed. We do not enter this debate as all of our judges are students.

[19]There are also no learning or time effects. Estimating the models using only the second half of the sample yields qualitatively the same results as the coefficients for the first and second half of the sample are not significantly different. Therefore judges do not seem to gain experience or suffer from fatigue over time. There is also no evidence for catch-up effects: judges do not classify significantly more (or less) individuals as liars in the first versus second half of displayed videos. These results are available on request.

exists within the group of individuals who declare honestly as well as within the group of individuals who underreported.

How can tax payers' individual dishonesty scores have implications for tax payers' choices whether to report truthfully or whether to underreport? If a tax payer is aware of how he or she is perceived by other people, then a tax payer who is perceived as an truthfully-looking person may be more inclined to underreport, and a less truthfully-looking tax payer may be more inclined to report truthfully. Of course, this logic implies that the tax payers make use of how they are perceived by others. Whether or not they do is generally difficult to measure.

We asked tax payers at the end of the tax compliance game to rate their own subjective probability for receiving an audit. Three answers were possible: higher than 50 percent, lower than 50 percent, and equal to 50 percent. Many aspects may affect this answer, including the actual experience in the experiment. Also, it is well-known that self-assessed data may be problematic. But notwithstanding these problems, we would expect a positive correlation between the dishonesty scores of subjects obtained received by the judges and their self-assessments about their subjective audit probabilities.

As defined above, the dishonesty score for each tax payer is equal to the fraction of judges who assess this particular tax payer as dishonest, i.e., the score is equal to zero if all judges assess a specific tax payer as honest and equal to one if all judges assess this tax payer as dishonest. Given this coding scheme, a positive correlation between the self-assessed audit probability and the dishonesty score would support hypothesis 2, i.e., it would suggest that the tax payers to some extent correctly assess how truthful they are perceived by others.

Figure 4 about here

Figure 4 shows the distribution of dishonesty scores for each of the three self-assessed audit probability categories. The boxplots clearly suggest a

positive correlation between the two variables. For example, the median dishonesty score within the first category (self-assessed audit probability smaller than 50%) is roughly equal to 0.42, while it is equal to 0.6 within the third category. We further summarize this positive correlation by running a linear regression. First, we create dummy variables for the three audit probability categories. Second, these dummy variables are entered as regressors to predict the average dishonesty score. The omitted reference category is "smaller than 50% ".

Table 2 about here

The results in Table 2 confirm the positive correlation found in the boxplots (Figure 4). The average dishonesty score for tax payers in category 1 (self-assessed audit probability smaller than 50%) is equal to 0.4258. The average dishonesty score for tax payers in category 2 (self-assessed audit probability equal to 50%) is roughly 7 percentage points higher compared to tax payers in category 1. Finally, the average dishonesty score for tax payers in category 3 (self-assessed audit probability above 50%) is roughly 14 percentage points higher compared to tax payers in category 1.

4.2. Evidence for self-selection

Having established that individuals differ in how they are perceived as honest or less honest, we are ready for hypothesis 3 to consider how these assessments affect tax payers' choice of their deception strategies We discussed why self-selection should make tax payers with a better (i.e., lower) dishonesty score more inclined to use deception strategies, and why this self-selection should be stronger for the high-penalty treatment conditions than for the low-penalty treatment conditions. The descriptive statistics indicate that this is indeed the case: the hit rate for underreporters is equal to 47.66 percent in the treatment with low fines and equal to 45.28 percent in the high fine treatment.

We now use our logistic mixed model (see the Table 1, equation (1) in column (1)) to check whether this model – which takes the heterogeneity of tax payers into account– generates the same findings as the descriptives. Evaluating the coefficients from this table yields predictions that are very close to the descriptives: the detection probability of an average liar is equal to $f(\hat{b}_0 + \hat{b}_2) = 47.70\%$ in the low penalty setup, whereas it is 44.97% in the set-up with high penalty. A likelihood-ratio-test indicates that the difference between these probabilities is marginally significant ($p < 0.1$). The individuals who choose a deception strategy in the high-penalty treatment are more successful than the individuals who choose to underreport in the low-penalty treatment. The liars in the high-penalty treatment are, on average, the better liars.

The difference between the groups is small compared to the variation in dishonesty scores modeled by the tax payer-video random intercept. However, from a theory point of view, it is surprising that there is an effect at all: sophisticated judges should essentially correct for the adverse incentives of individuals who have the "honest look" when they make audit choices, and when they make judgements about the likely truthfulness of a person. The data show that they may partially use such sophisticated judgements. Also, a close relationship between dishonesty scores and deception success requires that individuals' self-selection is based on their dishonesty scores. If they do not perfectly know how they are perceived by judges, or if the "honest look" is correlated with unobservable variables that make such individuals more averse towards using deception strategies, this can also weaken the relationship between dishonesty scores and the strength of self-selection.

We run two additional robustness checks regarding the selection hypothesis. First, we use only the observations that show clips with underreporting high-endowment tax payers and calculate for each judge two different hit rates: one for low penalty observations and the second one for the high penalty observations. This generates a paired data set, where each

of the $n = 231$ judges provides one pair of hitrates (x_{i1}, x_{i2}). A Wilcoxon signed rank test for paired data supports the conclusions from the mixed-effects model; the difference in the hitrates for liars is marginally significant $(p < 0.1)$.

Second, we use the same paired data set to fit a third mixed-effects model, in this case a linear mixed model. Thus, we use the 462 observations of the paired data set to estimate the equation

$$x_{it} = a_0 + a_1 \text{High Penalty} + u_i + \epsilon_{it}. \tag{2}$$

As before, *High Penalty* is a dummy indicating observations from the high penalty condition and u_i is a judge-specific random intercept. Table 3 compiles the results.

Table 3 about here

Note that the coefficients are normalized to reflect deviations from a deception detection rate of 50 percent. The results are in line with the previous results: The detection rate for high penalty videos is 2.38 percentage points smaller than the detection rate for the low penalty videos ($t = -1.69$, $p < 0.1$). Further, the small point estimate of σ_u^2 indicates that the variation of the hit rate due to unobserved judge-heterogeneity is very close to zero. Finally, it is worth noting that the hit rate for low penalty videos (47.66 percent) is significantly smaller than 50 percent ($p<0.05$). As the hit rate for high penalty videos is even smaller, the deviation from chance is significant as well.

4.3. Summary of results

In summary, there are two main findings: First, the data suggests that the videotaped tax payers have different detection probabilities. This holds for both honest and underreporting tax payers. As a side remark: the judges

in our experiment do not differ regarding their detection ability. The variation of tax payers' hit probability is large and is unrelated to observable characteristics like age or gender. Second, we find some evidence for a selection effect along the dimension of deceptive ability: The hit rate of judges is smaller for videos showing underreporters in the high-fine treatment than the respective hit rate for observations from the low-fine treatment. Questionnaire data on self-assessed audit probabilities further corroborates this evidence. The presence of self-selection may also explain the low overall hit rate of less than what could be obtained from pure random assessment.

5. Conclusion

This paper investigates experimentally the choice problems for which the attempt to deceive involves a material risk. Major motivational factors for this choice should be the benefit of successful deception and the cost if the deception attempt is detected, in comparison to the outcome in case of truth telling, and the likelihood for successful deception or detection. Individuals who feel confident about their deception abilities should, hence, be less likely to tell the truth. This reasoning suggests that choice and the self-selection implied is an important aspect for lie-catching in a natural compliance environment.

The insights from our experiment are threefold: First, and in line with hypothesis 1, we find major heterogeneity in the deception abilities among the individuals. Some individuals are poor liars and easily classified as deceivers, others are gifted in deceiving and hardly ever classified as liars. This classification of single individuals is consistent across judges. This consistency pattern holds for individuals who make deceptive statements as well as for individuals who make truthful statements. In contrast, we do not find heterogeneity among the judges' ability to detect deception. Second, subjects' self-assessed likelihood for an audit is positively related with their dishonesty score as conjectured by hypothesis 2. Hence, subjects can to some

extent correctly assess how truthful they are perceived by others. Third the self-assessed deception ability influences compliance choices. As a first hint, we find that liars whose deception is the outcome of their own choice are less frequently detected than in standard experiments where individuals are regularly forced to give a certain statement. In our data set, the overall hit rate of correctly classified statements is 47.35 percent. This deviation from pure chance is statistically significant. Moreover and in line with hypothesis 3, we find mild evidence for a selection effect: Tax payers who choose to underreport in a situation with high fines have lower dishonesty scores and are, therefore, perceived as more honest than the set of tax payers who choose to underreport in a situation with low fines. More precisely, the share of successfully detected deceptions drops from a 47.66 percent hit rate in the low-fine treatment to a 45.28 percent hit rate in the high-fine treatment. This is in line with the interpretation that, on average, individuals with stronger deception abilities choose to underreport if underreporting is more strongly discouraged by higher fines.

Our findings have implications for audit design, and the implications are not necessarily limited to tax compliance situations. The findings uncover a possible drawback of an audit mechanism that gives discretion to inspectors about who to select for an audit may result in poor results: discretionary decision making may discourage deception by weak liars more strongly, and it may discourage liars of superior deception ability less strongly. As a result, the set of individuals who do lie consists of individuals who have superior deception abilities. And this selection effect is stronger the more high-powered the incentives are: the higher the fines, the more individuals with low deception abilities are discouraged and the more capable deceivers are the individuals who choose to apply a deception strategy. This self-selection may lead to low detection rates - rates that even fall below the rates that can be achieved by a purely random audit.

Acknowledgements:. For providing laboratory resources we kindly thank MELESSA of the University of Munich and TU-Lab of the Technical University of Berlin. We thank Hans Müller for developing and programming the web-based environment. We thank Ray Rees and participants of the 3rd conference on "The Shadow Economy, Tax Evasion and Governance" at the University of Münster/Germany and of the Munich-Sydney-Conference on "The Law and Economics of Taxation" in Munich/Germany for helpful comments. The usual caveat applies.

References

[1] Allingham, Michael G., and Agnar Sandmo, 1972, Income tax evasion, a theoretical analyis, *Journal of Public Economics*, 1(3-4), 323-338.

[2] Andreoni, James, Brian Erard, and Jonathan Feinstein, 1998, Tax compliance, *Journal of Economic Literature,* 36(2), 818-860.

[3] Andreoni, James, and Ragan Petrie, 2008, Beauty, gender and stereotypes: Evidence from laboratory experiments, *Journal of Economic Psychology,* 29(1), 73-93.

[4] Barrick, Murray R., Jonathan A. Shaffer, and Sandra W. DeGrassi, 2009, What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance, *Journal of Applied Psychology*, 94(6), 1394-1411.

[5] Becker, Gary S., 1968, Crime and punishment: An economic approach, *Journal of Political Economy*, 76(2), 169-217.

[6] Bond, Charles F., and Bella M. DePaulo, 2006, Accuracy of deception judgments, *Personality and Social Psychology Review*, 10(3), 214-234.

[7] Brandts, Jordi, and Gary Charness, 2003, Truth or consequences: an experiment, *Management Science*, 49(1), 116-130.

[8] Brosig, Jeanette, 2002, Identifying cooperative behavior: some experimental results in a prisoner's dilemma game, *Journal of Economic Behavior & Organization*, 47(3), 275-290.

[9] Castillo, Marco, and Ragan Petrie, 2010, Discrimination in the lab: Does information trump appearance?, *Games and Economic Behavior,* 68(1), 50-59.

[10] Chander, Parkash, and Louis Wilde, 1992, Corruption in tax administration, *Journal of Public Economics*, 49(3), 333-349.

[11] Charness, Gary, and Martin Dufwenberg, 2006, Promises and partnership, *Econometrica,* 74(6), 1579-1601.

[12] Coricelli, Giorgio, Mateus Joffily, Claude Montmarquette, and Marie-Claire Villeval, 2010, Cheating, emotions, and rationality: an experiment on tax evasion, *Experimental Economics*, 13(2), 226-247.

[13] Crawford, Vincent, 2003, Lying for strategic advantage: rational and boundedly rational misrepresentation of intentions, *American Economic Review*, 93(1), 133-149.

[14] Darwin, Charles, 1872, The Expression of the Emotions in Man and Animals, London: John Murray, Albemarle Street.

[15] DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper, 2003, Cues to deception, *Psychological Bulletin,* 129, 74–118.

[16] DePaulo, Bella M., and Roger L. Pfeifer, 1986, On-the-job experience and skill at detecting deception, *Journal of Applied Social Psychology,* 16, 249–267.

[17] Dionne, Georges, Florence Giuliano, and Pierre Picard, 2009, Optimal auditing with scoring: theory and application to insurance fraud, *Management Science*, 55(1), 58-70.

[18] Dirks, Kurt T., Roy J. Lewicki, and Akbar Zaheer, 2009, Repairing relationships within and between organizations: building a conceptual foundation, *Academy of Management Review*, 34(1), 68-84.

[19] Eckel, Catherine C., and Ragan Petrie, 2011, Face value, *American Economic Review,* 101(4), 1497-1513.

[20] Ekman, Paul, and Wallace V. Friesen, 1974, Detecting Deception from Body or Face, *Journal of Personality and Social Psychology,* 29(3), 288–298.

[21] Erat, Sanjiv, and Uri Gneezy, 2012, White lies, *Management Science*, 58(4), 723-733.

[22] Fagart, Marie-Cécile, and Bernard Sinclair-Desgagné, 2007, Ranking contingent monitoring systems, *Management Science*, 53(9), 1501-1509.

[23] Frank, Robert H., Thomas Gilovich, and Dennis T. Regan, 1993, The evolution of one-shot cooperation: an experiment, *Ethology and Sociobiology*, 14(4), 247-256.

[24] Freud, Sigmund, 1959, Fragment of an Analysis of a Case of Hysteria. In: Collected papers, 3, New York: Basic Books.

[25] Frey, Bruno S., 1997, A constitution for knaves crowds out civic virtues, *The Economic Journal* 107, 1043-1053.

[26] Frey, Bruno S., and Benno Torgler, 2007, Tax morale and conditional cooperation, *Journal of Comparative Economics,* 35(1), 136-159.

[27] Gneezy, Uri, 2005, Deception: the role of consequences, *American Economic Review*, 95(1), 384-394.

[28] Greiner, Ben, 2004, An online recruitment system for economic experiments, University Library of Munich, MPRA Paper Nr. 13513.

[29] Habyarimana, James, Macartan Humphreys, Daniel Posner, and JeremyWeinstein, 2007, Why does ethnic diversity undermine public goods provision?, *American Political Science Review,* 101(4), 709-725.

[30] Hartner, Martina, Silvia Rechberger, Erich Kirchler, and Alfred Schabmann, 2008, Procedural fairness and tax compliance, *Economic Analysis and Policy,* 38(1), 137-152.

[31] Hartwig, Maria, and Charles F. Bond, 2011, Why do lie-catchers fail? A lens model meta-analysis of human lie judgments, *Psychological Bulletin,* 137(4), 643–659.

[32] Hindriks, Jean, Michael Keen and Abhinay Muthoo, 1999, Corruption, extortion and evasion, *Journal of Public Economics*, 74(3), 395-430.

[33] Holm, Hakan J., 2010, Truth and lie detection in bluffing, *Journal of Economic Behavior & Organization*, 76(2), 318-324.

[34] Holm, Hakan J., and Toshij Kawagoe, 2010, Face-to-face lying – An experimental study in Sweden and Japan, *Journal of Economic Psychology*, 31(3), 310-321.

[35] Kellerman, Barbara, 2006, When should a leader apologize and when not?, *Harvard Business Review*, 84(4), 72-81.

[36] Konrad, Kai A., Tim Lohse, and Salmai Qari, 2012, Compliance and the Subjective Audit Probability, *Max Planck Institute for Tax Law and Public Finance Working Paper*, 2011 – 18.

[37] Konrad, Kai A., Tim Lohse, and Salmai Qari, 2013, Dubious Versus Trustworthy Faces – What Difference Does it Make for Tax Compliance?, *CESifo Working Paper,* 4373.

[38] Konrad, Kai A., and Salmai Qari, 2012, The last refuge of a scoundrel? Patriotism and tax compliance, *Economica*, 79(315), 516-533.

[39] Ksh, Jhaljit Singh, 2008, On tax evaders and corrupt auditors, *Journal of International Trade & Economic Deveolpment*, 17(1), 37-67.

[40] Lombroso, Cesare, 1876, L'uomo delinquent. In rapporto all'antropologia, alla giurisprudenza ed alle discipline carcerarie, Turin: Bocca.

[41] Lundquist, Tobias, Tore Ellingsen, Erik Gribbe, and Magnus Johannesson, 2009, The aversion to lying, *Journal of Economic Behavior & Organization*, 70(1-2), 81-92.

[42] Mobius, Markus. M., and Tanya S. Rosenblat, 2006, Why beauty matters, *American Economic Review,* 96(1), 222-235.

[43] Ockenfels, Axel, and Selten, Reinhard, 2000, An experiment on the hypothesis of involunatary truth-signalling in bargaining, *Games and Economic Behavior*, 33(1), 90-116.

[44] Reinganum, Jennifer F., Louis L. Wilde, 1985, Income tax compliance in a principal-agent framework, *Journal of Public Economics*, 26(1), 1-18.

[45] Rosaz, Julie, and Marie Claire Villeval, 2012, Lies and biased evaluation: A real-effort experiment, *Journal of Economic Behavior & Organization*, 84(2), 537-549.

[46] Sánchez-Pagés, Santiago, and Marc Vorsatz, 2007, An experimental study of truth-telling in a sender-receiver game, *Games and Economic Behavior*, 61(1), 86-112.

[47] Shapiro, Carl, and Joseph E. Stiglitz, 1984, Equilibrium unemployment as a worker discipline device, *American Economic Review*, 74(3), 433-444.

[48] Slemrod, Joel, 2007, Cheating ourselves: The economics of tax evasion, *Journal of Economic Perspectives*, 21(1), 25-48.

[49] Solnick, Sara J., and Maurice E. Schweitzer, 1999, The influence of physical attractiveness and gender on ultimatum game decisions, *Organizational Behavior and Human Decision Processes,* 79(3), 199-215.

[50] Torgler, Benno, 2006, The importance of faith: Tax morale and religiosity, *Journal of Economic Behavior & Organization*, 61(1), 81-109.

[51] Vanberg, Christoph, 2008, Why do people keep their promises? An experimental test of two explanations, *Econometrica*, 76(6), 1467-1480.

[52] Vrij, Albert, 2008, *Detecting Lies and Deceit*, 2nd ed. Chichester: Wiley.

[53] Wilson, Rick K., and Catherine C. Eckel, 2006, Judging a book by its cover: Beauty and expectations in the trust game, *Political Research Quarterly,* 59(2), 189-202.

[54] Yim, Andrew, 2009, Efficient committed budget for implementing target audit probability for many inspectees, *Management Science*, 55(12), 2000-2018.

6 Figures and Tables

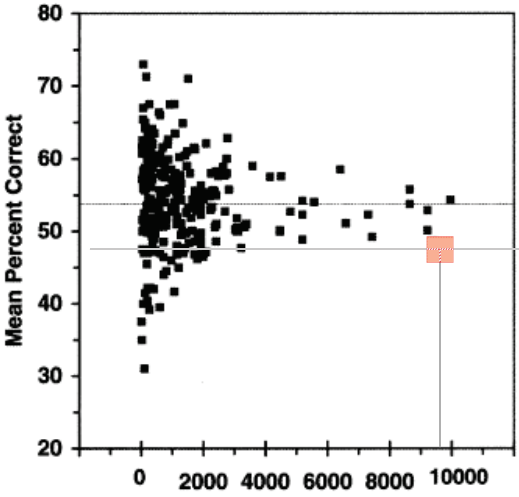

Figure 1: Percentage of correct judgements and number of observations in our experiment compared to findings in previous lie-catching studies. The center of the red square locates the outcome of our experiment. Source: Bond and Depaulo (2006, p.222) and own calculations.

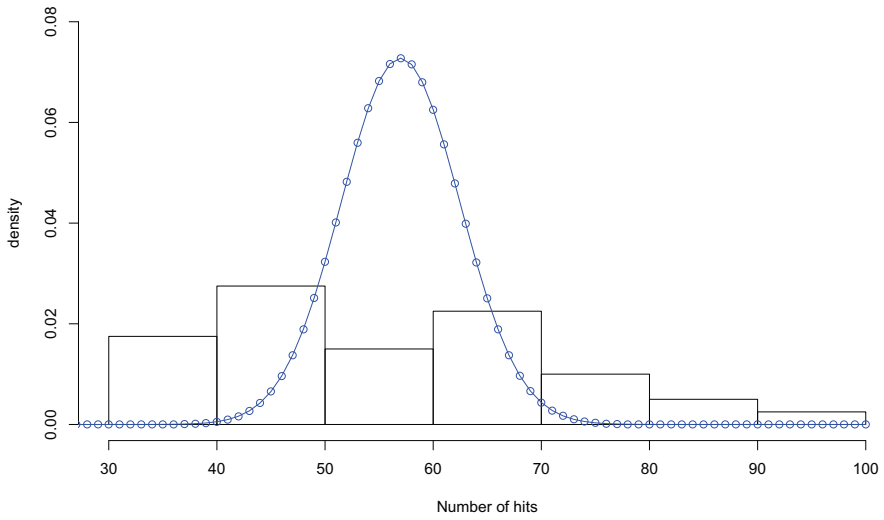

Figure 2a: Distribution of hits for taxpayers, Sample 1 (120 assessments). Each taxpayer in sample 1 was assessed by 120 different judges. If the detection probability would be constant for all taxpayers and equal to the average detection rate (47.35%), the distribution would follow a binomial distribution with $p = 0.4735$ and $N = 120$ (blue curve). The observed distribution is seemingly not compatible with this binomial distribution.

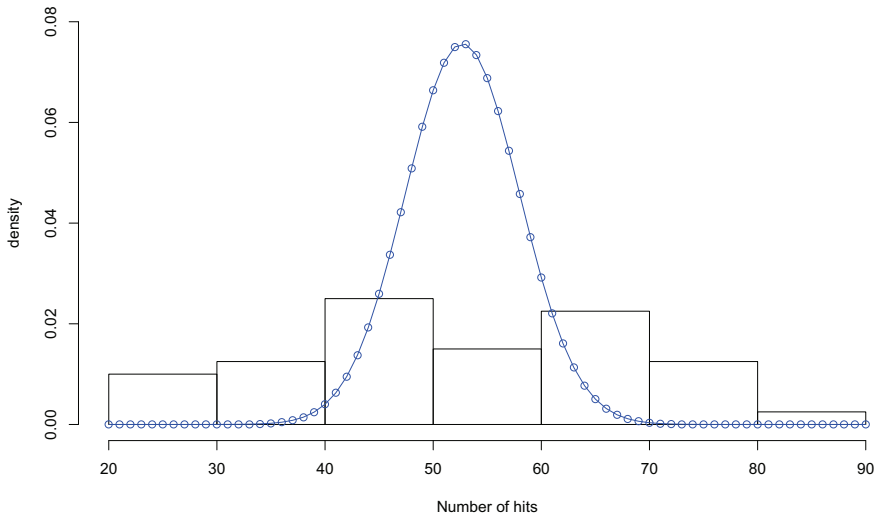

Figure 2b: Distribution of hits for videos, Sample 2 (111 assessments). Each taxpayer in sample 2 was assessed by 111 different judges. If the detection probability would be constant for all taxpayers and equal to the average detection rate (47.35%), the distribution would follow a binomial distribution with $p = 0.4735$ and $N = 111$ (blue curve). The observed distribution is seemingly not compatible with this binomial distribution.

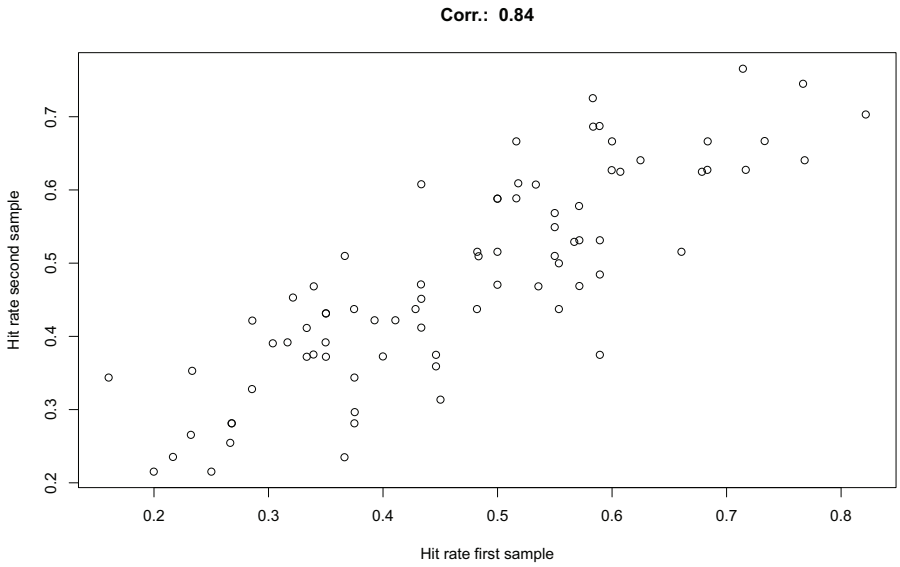

Figure 3: Correlation across judges for single tax payers. Each entry represents the hit ratios m_1/n and m_2/n for each of the single tax payers, where m_1 is the hit rate of the first half of judges assessing this tax payer and m_2 is the hit rate of the second half of the judges assessing this taxpayer, and n is the total number of judges who assessed this tax payer.

	(GLMM 1)	(GLMM 2)
Constant	−0.2681*	−0.6099
	(0.1261)	(1.7718)
	[0.0335]	[0.7307]
High Penalty	0.3782*	0.3737*
	(0.1782)	(0.1784)
	[0.0338]	[0.0362]
Liar	0.1761	0.1852
	(0.1784)	(0.1831)
	[0.3234]	[0.3119]
High Penalty × Liar	−0.4880•	−0.5474*
	(0.2521)	(0.2704)
	[0.0529]	[0.0429]
Female		0.0295
		(0.2089)
		[0.8877]
Age		−0.0275
		(0.0240)
		[0.2511]
Height		0.0025
		(0.0094)
		[0.7949]
Protestant		−0.0616
		(0.1924)
		[0.7487]
Other christian		0.3612
		(0.2591)
		[0.1633]
Non-christian		0.6156
		(0.5745)
		[0.2840]
No religious denomination		0.1150
		(0.1836)
		[0.5310]
Religious (yes)		−0.0833
		(0.1775)
		[0.6389]
Church visit: 1-5 times		0.2544
		(0.1640)
		[0.1208]
Church visit: >5 times		0.2378
		(0.2879)
		[0.4088]
Difficult to lie: yes		−0.1233
		(0.1611)
		[0.4441]
Gambling: yes		−0.1162
		(0.1669)
		[0.4861]
σ_u	0.0000	0.0000
σ_v	0.2807	0.2489
Log-likelihood	−6159.9826	−6155.8412
AIC	12331.9653	12347.6824
N	9240	9240

Standard errors in round parentheses, p-values in brackets

•p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 1: Generalized (logistic) mixed effects model without and with socio-economic variables as controls. The models predict the hit probability for a tax payer-video as a function of the treatment conditions (column 2 includes additional controls) and random effects for tax payer-videos and judges. The estimated standard deviation of the tax payer random effects ($\sigma_v = 0.28$) indicates a considerable degree of heterogeneity regarding the tax payer-specific hit probabilities.

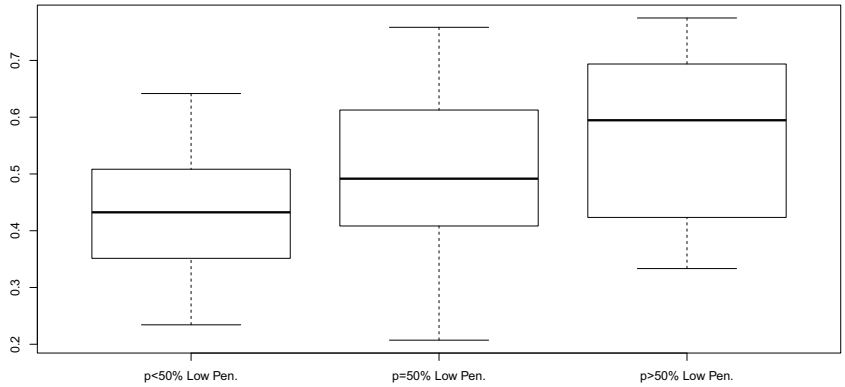

Figure 4: Correlation of self-assessed audit probability and assessment/dishonesty-score by judges. The dishonesty score for each tax payer is equal to the fraction of assessors/judges who assess the video clip of this particular subject as dishonest, i.e., the score is equal to zero if all assessors/judges assess the video of this subject as honest and equal to one if all assessors/judges assess this subject as dishonest. The videotaped tax payers were asked to assess their subjective audit probability according to three categories (<50%, =50%, >50%). The boxplots indicate a positive correlation between the self-assessed audit probabilities and the dishonesty score derived from the judges' assessments.

Constant	0.4258***
	(0.0242)
equal to 50 percent	0.0753*
	(0.0342)
above to 50 percent	0.1394***
	(0.0368)
R-squared	0.1592
F	7.2917
p	0.0013
N	80

Standard errors in parentheses

•p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 2: Linear Model / Regression of Dishonesty Score on indicator variables for the three self-assessed audit probability categories. See the boxplot (Figure 4) for a description of the variables. The regression confirms the positive correlation between self-assessed audit probability and the dishonesty score derived from the judges' assessments found in the boxplots (Figure 4).

Constant	−0.0234*
	(0.0101)
High Penalty	−0.0238•
	(0.0141)
σ_u^2	0.0005
Log-likelihood	204.5750
AIC	−401.1500
N	462

Standard errors in parentheses

•p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 3: The table shows the estimation results from a linear mixed effects model. The explained variable is the hit rate. The negative coefficient of the treatment variable "High Penalty" is in line with the self-selection hypothesis according to which the set of underreporters in the high-penalty treatment consists of tax payers with superior deception abilities.