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

Working Paper

Compliance with endogenous audit probabilities

DIW Discussion Papers, No. 1493

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Konrad, Kai A.; Lohse, Tim; Qari, Salmai (2015) : Compliance with endogenous audit probabilities, DIW Discussion Papers, No. 1493, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/112277

DIW BERLIN

1493

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2015

Compliance with Endogenous Audit Probabilities

Kai A. Konrad, Tim Lohse and Salmai Qari

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Compliance with endogenous audit probabilities*

Kai A. Konrad† Tim Lohse‡ Salmai Qari§

July 9, 2015

This paper studies the effect of endogenous audit probabilities on reporting behavior in a face-to-face compliance situation such as at customs. In an experimental setting in which underreporting has a higher expected payoff than truthful reporting we find an increase in compliance of about 80% if subjects have reason to believe that their behavior towards an officer influences their endogenous audit probability. Higher compliance is driven by considerations about how own appearance and performance affect their audit probability, rather than by social and psychological effects of face-to-face contact.

JEL-Codes: H26, H31, C91, K42

Keywords: Compliance, audit probability, tax evasion, face value, customs

1 Introduction

The audit probability shows up in the theory of crime (Becker 1968) as a key determinant of deceptive behavior. This probability can be exogenously given; but in many real life instances it is not. When individuals communicate face-to-face with tax authorities, with their boss about budgeting costs of a business project they are responsible for or with their professor about potential plagiarism of their thesis etc., they have discretion as regards their behavior. In these situations individuals need to conjecture how their appearance and performance affect their *endogenous* probability of receiving an audit. Some individuals may be self-confident in a face-to-face situation since they believe that they are highly capable of deception. Therefore they assume a low probability for themselves. Others may believe that they are prone to bad luck and that Fortuna will always pick on them whenever they attempt to cheat. This paper makes a first attempt to understand the effect of endogenous audit probabilities on compliance behavior in a face-to-face situation.

*For providing laboratory resources we kindly thank MELESSA of the University of Munich. This article was finished while the third author was a visiting researcher at the DIW Berlin, Department of Public Economics and he is grateful for their hospitality.

†Max Planck Institute for Tax Law and Public Finance, Marstallplatz 1, 80539 Munich, Germany, Email: kai.konrad@tax.mpg.de

‡Berlin School of Economics and Law, Badensche Straße 52, 10825 Berlin, Germany, and Max Planck Institute for Tax Law and Public Finance, Marstallplatz 1, 80539 Munich, Germany. Email: tim.lohse@hwr-berlin.de (corresponding author)

§Berlin School of Economics and Law, Badensche Straße 52, 10825 Berlin, Germany, Email: salmai.qari@hwr-berlin.de

The experimental tax compliance framework we use resembles the situation at customs. We analyze three treatments to study whether or not an individual's compliance decision is affected (at all) by the fact that her audit probability is endogenous (rather than exogenous) and, therefore, depends on her behavior. In our main treatment (later on labeled T3) individuals meet face-to-face with an officer to whom they must make an oral compliance declaration and who assesses the honesty of their declaration. This assessment is influential for the question whether or not the individual receives an audit. This implies that the audit probability is endogenous. We compare this treatment T3 with a scenario in which there is as little room as possible for subjective deviations from an exogenously imposed audit probability. In that treatment (labeled T1) the audit probability is hard wired in a computerized audit system and declarations occur directly via a computer. Hence, there are no officers the individuals meet or could interact with. This benchmark treatment T1 resembles the prototypical set-up of a large set of computerized tax compliance experiments used throughout the literature.

However, a whole set of aspects of the compliance situation changes between a purely computerized audit system with an exogenous audit probability as in treatment T1 and a compliance situation with face-to-face contact with an officer and an endogenous audit probability as in treatment T3 (see, e.g., Holm and Kawagoe 2010). Personal contact with a customs officer (rather than declaring on a computer screen) can, for instance, cause a higher mental cost of lying (Vanberg 2008, Lundquist, Ellingsen, Gribbe and Johannesson 2009), may invoke shame (Coricelli, Joffily, Montmarquette and Villeval 2010) or guilt aversion (Charness and Dufwenberg 2006, among others) or trigger other psychological effects on self-image etc. These are relevant dimensions for deception decisions, and it is evident from some of these studies that the strength of the psychological motivations depends on the specific environment of the compliance situation.[1] In order to take into account these traces of human psychology that may come along with a face-to-face compliance situation, we design an additional control treatment (labeled T2). Treatment T2, on the one hand, resembles T3 as the individuals also meet a customs officer in person, make an oral declaration etc, but, on the other hand, follows T1 since the same exogenous audit rules with a fixed and known audit probability are applied. All psychological effects from face-to-face contact with an officer and of the specific compliance situation are present in treatment T2. But unlike in T3 there is no room for the considerations about how own behavior affects the audit probability as this probability is exogenously fixed. This control treatment T2 takes stock of the full set of purely psychological effects of face-to-face reporting. What remains in the comparison between this control treatment and our main treatment T3 is how considerations about own appearance and performance in a face-to-face

[1] Coricelli et al. (2010) measure the emotional arousal of cheaters if their picture is displayed publicly. See also Feldman and Slemrod (2007) and especially Kleven, Knudsen, Kreiner, Pedersen and Saez (2011), who analyze data from Danish tax authorities and find evidence that tax evasion is higher for self-reported income.

situation affect the endogenous audit probability and consequently compliance behavior.[2] This motivates the title of the paper.

The literature on tax evasion has identified the audit probability as a crucial determinant for compliance (see, e.g. Allingham and Sandmo 1972, Yitzhaki 1974, Reinganum and Wilde 1985, 1986, and Chander and Wilde 1998).[3] Our research contributes to this literature and especially to the experimental literature on tax compliance. Alm (2010, p. 645-47) reports that a large share of these experimental efforts concentrated on variables such as the probability of an audit and the size of a penalty in case of misreporting that are each related to the material-rewards-oriented decision models on tax compliance. However, in the literature the audit probability is usually exogenously determined. In contrast, this paper considers the possible role of subjects' perceptions about whether or not they can influence the beliefs of others concerning their honesty in a compliance situation. Much of the literature assumes away the complexities of tax declarations (e.g. the uncertainty about how certain declaration choices and tax planning activities would be viewed as legal or illegal by the tax office and tax authorities) and reduces the problem to a compliance decision similar to the decision of travelers at customs.[4] In this sense, we address tax evasion in the framework of a customs compliance situation.[5]

Our research also complements and extends the findings about the value of a face by Eckel and Petrie (2011). They run a trust game experiment in which individuals are allowed to buy a photo of their counterpart before taking an economic decision. From such a photo individuals may infer some characteristics that have been identified to play an important role.[6] Eckel

[2]To illustrate the point, think of customs control with customs officers watching a traveler, potentially talking to this traveler, and then deciding on whether to inspect this person more closely. Many individuals may think that how they dress, how they look, whether they seek or avoid eye contact with the customs officers, whether they have a straight look or whether their eyes move unsteadily, and whether they talk eloquently with a confident or with a broken voice etc., may affect the decision about whether or not they are chosen for an audit. Even if travelers know that the share of travelers who cheat may be anticipated and the share of travelers who can be chosen for an audit is essentially given by the capacity of customs officers and the overall number of travelers, they need not think that this average quota is the probability that applies to them individually.

[3]Besides the audit probability, other determinants may - and have been identified to - affect individuals' decisions about truthful compliance. These include intrinsic motivation (Frey 1997), an inclination for pro-social behavior (Frey and Torgler 2007) or pro-ethical behavior (Boadway, Marceau and Mongrain 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.

[4]Theory contributions focusing on customs compliance more explicitly are Thursby, Jensen and Thursby (1991) and Yaniv (2010).

[5]The underlying theory of our experimental analysis also has some links with the theory of beliefs about beliefs. As we study a customs compliance situation in an environment that uses elements of a field experiment (see section 2) implying that the procedure is very time-consuming, we focus on subjects' actions as an outcome variable. This approach of studying beliefs is in line with work by e.g. Weizsäcker (2003) or Camerer, Ho and Chong (2004). See Manski and Neri (2011) and Costa-Gomes and Weizsäcker (2008) for examples and discussion of more complex approaches of eliciting beliefs.

[6]Such characteristics are, e.g., beauty (Mobius and Rosenblat 2006, Wilson and Eckel 2006), ethnicity (Habyarimana, Humphreys, Posner and Weinstein 2007), gender (Solnick and Schweitzer 1999, Andreoni and Petrie

and Petrie (2011) find the informational value of a face to be non-zero and observe a change in economic behavior once the veil of anonymity is lifted. Our analysis sheds light on the crucial question what exactly makes individuals willing to pay for unveiling their counterpart's face. Two possible explanations come to mind: First a pure face-effect: individuals do not like anonymity per se and are consequently willing to remove this uncertainty about their counterpart's face. Second a face-effect in combination with the discretionary consequences by their counterpart: unveiling the counterpart's face is only valuable when individuals really interact with each other, i.e. in a situation in which a counterpart's decision affects oneself. With our experimental design we are able to disentangle these two explanations: Moving from treatment T1 to T2 removes anonymity and moving from T2 to T3 allows additionally for economic interaction. Hence, a significant change in compliance behavior between T1 and T2 could be an indication for a pure face-effect. In contrast, a change in compliance behavior when moving from T2 to T3 would indicate that it is not a face-effect alone but also the discretionary consequences by the counterpart that make unveiling the face valuable.

The main findings from our experiment are as follows: There is evidence that the endogenous audit probability alters reporting behavior. In an environment in which subjects are induced to cheat (i.e., in which the expected monetary payoff from cheating is positive), the proportion of subjects who report truthfully in the main treatment T3 is higher if the audit probability depends on the officer's decision - even for the same given overall audit quota for cheaters. The share of subjects who truthfully declare increases roughly by 80% compared to declaration behavior in the computerized benchmark treatment with a purely random audit. In contrast, there is no difference in the declaration behavior between the benchmark treatment T1 and the control treatment for psychological effects, T2. Consequently, the treatment effects are not generated by the change from a fully computerized set-up T1 to face-to-face declaration in front of a human customs officer, T2. The social and psychological effects of personal contact with a customs officer as captured by the control treatment are insignificant: either the different psychological effects cancel each other out or the effects are simply very weak. The increased compliance in treatment T3 can, therefore, be attributed to the fact that the audit probability is endogenous. Moreover, with respect to Eckel's and Petrie's observation of the value of a face, our findings indicate that individuals are more inclined to attribute a certain value to a face when they expect to interact with that counterpart.

The remainder of the paper is organized as follows: in Section 2 we explain the experimental design and derive the testable hypotheses. Section 3 shows our findings about the role of the endogenous audit probability. Section 4 inquires as to the robustness of the estimated effects and Section 5 concludes.

2008) or race (Castillo and Petrie 2010).

2 Evidence, theory and the experimental setup

Insights from real world customs

In order to get a better understanding of what really happens in a face-to-face compliance situation, and before designing the laboratory experiment, we conducted 17 in-depth interviews with customs officers at three different German international airports.[7] We were interested in the perceptions of customs officers, particularly regarding their view about the effectiveness of their own audit behavior and how they affect travelers' compliance behavior. In the following, we explain five key findings from the interviews, which were conducted in 2009 and 2010, and their importance for our experimental design. (1) Sixteen out of seventeen officers agreed that compliance behavior is affected if travelers see customs officers at the customs gate, compared to a situation in which they do not see any customs officer standing nearby. Officers reported people getting nervous or turning red when just being watched by an officer. Concerning the case in which a traveler is chosen for inspection, about half of the officers stated that they can tell right away whether or not the chosen traveler tried to smuggle goods. (2) The perception was prevalent that the mere presence of customs officers induces more honest behavior. Two thirds of the officers reported that their physical presence increased the share of travelers who chose to declare something. This inspired us to design the experiment in such a way that we can separate the effects of pure face-to-face contact from the effects which come along with an endogenous audit probability: in the first case, travelers may change their behavior because they suffer from lying to an individual, for example. In the latter case, however, they change their behavior strategically as they make use of the fact that they can influence their audit probability through their behavior. (3) We asked officers to rank three alternative institutional frameworks according to which framework most effectively induces honest behavior. These included the voluntary self-selection of travelers into an exit path for travelers who have nothing to declare and an exit path for travelers who have to make a declaration, a written and signed declaration form, and verbal customs declarations face-to-face with an officer. About half of the officers interviewed considered declarations with officers being there in person as the most effective scenario. This suggests that the interviewed customs officers expect subjects to exist whose endogenous audit probability is rather high in a face-to-face compliance situation, i.e. who believe themselves to be bad liars. (4) All customs officers unambiguously agreed that there are also travelers who are very good at lying, i.e. that have rather low endogenous audit probabilities. Our experiment shall examine whether these expectations about the existence of individuals with high and low endogenous audit probabilities are justified. (5) From officers'

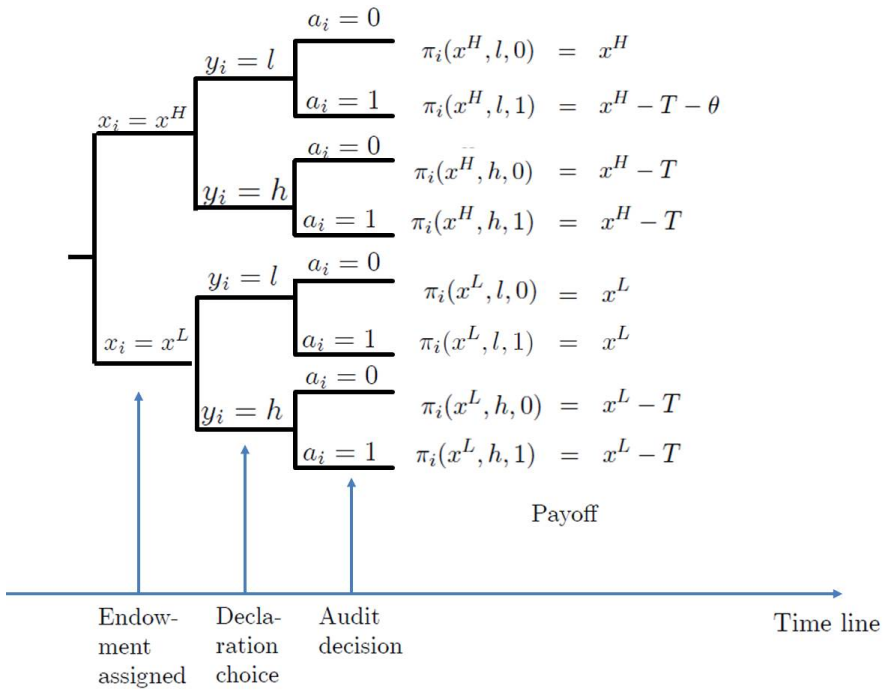
[7] The customs officers were encouraged by the customs administration to volunteer for these interviews. The 28 interview questions were raised by an interviewer and the (semi-open) answers were voice-recorded and transcribed. These also included questions about the officer's age, experience and position.

answers, we can conclude that customs officers have a large degree of freedom in their choices about whom to audit and develop their own heuristics in their job. This is important as it suggests that belief formation about one's endogenous audit probability has a legitimate place in real compliance contexts; travelers may rightly believe that their characteristics and behavior affect their individual probability of being audited.

A simple model of compliance

We consider the following standard compliance problem resembling the situation at customs. An individual i has an endowment $x_i \in \{x^L, x^H\}$ with $x^L < x^H$. This endowment is a random draw from a probability distribution that is common knowledge. The probability for a low endowment (x^L) is 0.2 and the probability for a high endowment (x^H) is 0.8. Each individual knows her own endowment. Customs cannot observe i's endowment directly, but knows the probability distribution. At customs i must declare her own endowment and chooses between two possible reports: $y_i \in \{h, l\}$. Customs receives this compliance report. This report is followed by a process that either leads to an audit ($a_i = 1$) or not ($a_i = 0$). The audit perfectly reveals the individual's true endowment. Figure 1 sketches the time line of the actions and shows the payoffs for individual i which are (exogenously) linked to the different combinations of endowments and actions.

Figure 1: The compliance model



These payoffs conform with the intuitive outcomes: low-endowment individuals pay no duties, if they report truthfully, regardless of whether they receive an audit or not. If an individual reports a high endowment, the individual has to pay a duty equal to T, regardless whether an audit occurs or not. Individuals with a high endowment who report truthfully have to pay a duty equal to T, regardless of whether they receive an audit. Individuals with a high endowment who declare a low endowment receive different payoffs dependent on whether they receive an audit or not. If an individual declares l and is not audited, neither a duty nor a fine is to be paid. If a high-endowment individual who reports a low endowment receives an audit, the individual has to pay the duty $T > 0$, and, in addition, a surtax that is equal to $\theta > 0$.

Given this set-up, we can safely assume that low-endowment individuals who maximize their payoffs report truthfully. Therefore, in the following, our focus is set on individuals with high endowments – the related variables are denoted with a superscript H – and their compliance problem.[8] A high-endowment individual i, who maximizes her own expected monetary payoff, prefers to report truthfully, if

$$x^H - T > p_i^H(x^H - T - \theta) + (1 - p_i^H)x^H, \tag{1}$$

where p_i^H is the probability that i, endowed with x^H, attributes to being audited in the case in which she declares l. In other words, p_i^H is the probability of being caught when cheating. An individual who cares about monetary incentives should, therefore, only be indifferent between compliance and non-compliance if

$$T = \frac{p_i^H}{1 - p_i^H}\theta. \tag{2}$$

Much of the further analysis is affected by how p_i^H is determined: p_i^H is exogenous in T1 and T2, but endogenous in T3.

Procedure of the experiment

We conducted the experiment in the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) in 2011. Twelve sessions took place. The participating subjects were students of diverse fields at the University of Munich. [9] In each session there were 20 subjects, totaling a number of 240 subjects, who participated in the role as travelers. Each session

[8] We assume all individuals to be risk neutral. Rabin (2000) shows that, within the expected-utility framework, anything but risk neutrality over modest stakes would imply rather unrealistic risk aversion over large stakes. In our empirical analysis we generate a risk measure by using data from a standard risk elicitation game in the style of Holt and Laury (2002) which participating subjects had to play. This risk measure has only little explanatory power in our data (see section 4 below for details).

[9] The participants were recruited from the subject pool of MELESSA, which mainly consists of students of a variety of fields. The software ORSEE (Greiner 2004) was used.

started with a reading of the instructions, which were also distributed in written form. Then the subjects had to go through an introductory computerized quiz, which took them about 10 minutes. The quiz outlined each possible payoff situation of the upcoming experiment. The participants had to calculate the resulting monetary payoff in each of the situations. They could only move on with the quiz if their answer was correct. Through this we ensured that the participants fully understood the nature of the experiment, especially the audit mechanism. Then the actual experiment started. Each individual participated in only one of the treatments. Participants in T1 played the compliance game for ten independent rounds. Since each round in treatments T2 and T3 takes more time than a round in T1, participants in T2 and in T3 played exactly four rounds. Between each round, the person who served as customs officer was replaced such that each participant met each of the four officers only once. After the experiment, the subjects had to answer an exit questionnaire asking them for their gender, their field of study, etc. This was followed by a standard risk elicitation game in the style of Holt and Laury (2002). Participants had to compare ten pairs of lotteries sequentially. This game also took about ten minutes and the participants were able to generate additional income, since, in the end, the computer selected one of the ten situations and simulated the chosen lottery. Average income from performing this task was 2.20 EUR. After all of that, participants received their earnings from the experiment, plus the outcome of that lottery, plus a show-up fee of 4 EUR. Total average earning was 16.6 EUR. The participants' earnings from the experiment consisted of the outcome of one specific round, which was randomly drawn by the computer. The currency in the laboratory was named talers.[10] The four customs officers, in line with reality, were paid flat.[11]

The three treatments of the experiment

Our experiment addresses the role of individuals' assessments about their own ability to influence customs officers. Inspired by the insights from our survey among real customs officers, we consider three different treatments, labeled T1, T2, and T3, and apply a between-subject design.[12] Throughout all treatments, care is taken that the participating subjects remain anonymous (or at least pseudonymous in T2 and T3) and do not exchange views or learn about other subjects' monetary payoffs either during the experiment or at the end of the ex-

[10] In the treatment T1, 1000 talers were converted into 10 EUR. In T2 and T3, 1000 talers were converted into 16 EUR. With these different exchange rates, we ensured that the participants' expected payoffs per unit of time they contributed were the same if they showed the same choice behavior in all treatments since the sessions with treatments T2 and T3 lasted longer.

[11] A further incentivation of the officers is not necessary since the research focus is entirely on the declaring individuals. Besides, recall that each customs officer encounters each subject just once.

[12] For a detailed discussion about a between- versus a within-subject design of an experiment see Charness, Gneezy and Kuhn (2012).

periment when payments are made. In all treatments, at the beginning of each round, an individual i learns the value of her endowment while sitting in front of a computer. The value is private information and it is equal to $x^H = 1000$ if it is high and $x^L = 400$ if it is low, with respective probabilities 0.8 and 0.2. If an individual reports $y_i = h$, she has to pay customs duties of $T = 200$. If an individual with a high endowment reports l and is audited, the true value x^H is found. Then the subject has to pay customs duties of $T = 200$ plus a fine θ. If an audit occurs, it is carried out by the computer. The individual gets to know at the end of each round whether she was audited or not.[13] The three treatments T1-T3 differ in how the declaration is made and who is selected for an audit.

In the fully computerized baseline treatment (T1), the individual is asked on the computer screen to declare a high or a low endowment: $y_i \in \{h, l\}$. The high-endowment individual knows (as this is written down as part of the instructions) that the computer chooses $p_i^H \equiv 0.5$ if $y_i^H = l$. The subjects make these decisions in a laboratory room in which 20 subjects perform the same task independently. As $p_i^H \equiv 0.5$ is given exogenously, each subject's task is formally independent of the tasks and choices of other individuals, i.e. who ever decides to cheat faces a 50% chance of being caught.[14] In T1, for $p_i^H \equiv 0.5$, the indifference condition (2) reduces to $T = \theta$. The own-material-interest prediction is that a high-endowment subject should choose $y_i^H = l$ if $T > \theta$ and $y_i^H = h$ if $T < \theta$.

In the second treatment (T2), the subjects again first learn their endowment while sitting in front of a computer in the same laboratory room as in T1. The 20 subjects wait until they are asked sequentially to walk into one of two separate neighboring rooms. The sequence of the subjects is determined randomly.[15] In each of the two rooms a person in the role of a customs officer waits for subjects and sees, on a computer screen, the identification number of the subject entering.[16] The customs officer first confirms the entrant's identification number. Then, the

[13] Alm, Cherry, Jones and McKee (2010) object to what they call the traditional enforcement paradigm which considers taxpayers as "potential criminals" (p. 577). They argue that it is rather the complexity and unclarity of tax schedules that leads to an unintentionally high degree of tax evasion. Our setting focuses solely on self-reported income with an individual declaration situation that is rather easy. Therefore, we are able to rule out complexity or unclarity as an explanatory factor for dishonest behavior.

[14] Beside the audit mechanism used here – audit of 50% of all high-endowment individuals with $y_i^H = l$ – an alternative mechanism could be an audit of 50% of all individuals declaring $y_i = l$, regardless of their true endowment. However, from the point of view of a potential cheater, both mechanisms work in exactly the same way. The selection of a particular mechanism is therefore solely a framing issue. We believe the audit mechanism used here is more useful for an experimental study as it is easiest to understand.

[15] By design the experiment precluded subjects from meeting each other during the compliance procedure or from inferring who preceded or succeeded them in their room. All individuals not currently active in complying saw the request "please wait" on their screen in the room with the terminals in the laboratory room. This room has several doors. People were asked to leave the room and return via separate doors. Along with the instructions, participants had received a map showing the position of the two rooms in question. Signs also directed them to their respective rooms.

[16] To facilitate the comparison of the compliance behavior across treatments, the communication between the customs officer and the subjects has to be standardized. More importantly, it is necessary to ensure anonymity for the subjects. For these reasons, customs officers were not recruited from the pool of stu-

subject has to report y_i, i.e., whether he or she has something to declare. The officer enters the subject's report in the computer. The subject returns to the laboratory room. The first round is over after ten subjects have reported to one officer and the other ten subjects to the other officer. The officer in one room is female, the other officer is male, in each round. Subjects meet each of the four officers only once.[17] The audit mechanism is the same as in T1: in T2 the high-endowment subject also knows (as this is written in the instructions) that the probability of being caught when cheating is $p_i^H \equiv 0.5$, exogenously given and, hence, independent of the subject's (or other subjects') general appearance or performance. The customs officer has no active decision role. This treatment takes into account that subjects having to report face-to-face to a real person rather than to a machine may make a difference, as their subjective cost of misreporting need not be the same in both situations.[18] Differences between T1 and T2 can be attributed to the psychology of the compliance decision, such as attitudes toward lying in personal communication, guilt, shame etc., that may be caused by the procedural and situational difference between T1 and T2. We conduct the control treatment T2 in order to control for the sum of these psychological effects when determining the effect of endogenous audit probabilities on compliance behavior. In the following, we refer to these psychological effects as the pure face-to-face effect and distinguish it from the effect of probability formation which comes into play in the third treatment (T3).

This treatment (T3) is similar in structure to treatment T2, but the crucial difference is how p_i^H is determined. In T3 the audit probability p_i^H is not exogenous anymore as in T1 and T2, but endogenous: the customs officer ranks the subjects according to his assessments about the honesty of their declarations.[19] This ranking is used to determine who receives an audit: Out of all high-endowment subjects who declared $y_i^H = l$, the more suspicious half according to the officers ranking is audited. This implies that in the aggregate, precisely half of all cheaters receive an audit. The aggregate audit rate (0.5) is constant across the three treatments for all potential cheaters and the same as in T1 and T2, thereby removing a potential confounder of the treatment effects. These rules are common knowledge as this is also what is explained to them in the instructions. In T3, subjects may be wondering how their appearance and performance affect the assessment of the customs officer of them being

dent subjects. Instead, young employees and contract workers from the Max Planck Institute played the role of customs officers. Travelers were not informed about the officers' flat payment scheme. They had no information about whether the officers were 'amateurs' in this role.

[17] The data analysis shows that there are no 'officer effects', i.e. subjects' reporting behavior does not vary by officer or the officer's gender. For a more detailed gender analysis see Lohse and Qari (2014).

[18] See, for instance, Lundquist et al. (2009) for some evidence that individuals dislike lying, particularly in free-form communication. Also, work by Coricelli et al. (2010) suggests that emotional cost of cheaters being caught is higher if cheating is made public.

[19] Grades ranged from 1 (=very credible) to 10 (=not credible at all). The audit mechanism made use of these grades.

honest or not. A subject i considers her endogenous audit probability p_i^H to be lower than 0.5 if she expects to be more likely to be assessed as an honest traveler than the median among all other subjects who choose to underreport. The rather low endogenous audit probability of such an individual leads her to comply less truthfully in T3 than in T1 and T2. We call such individuals 'strong liars'. Sufficiently strong liars would declare truthfully in T1 or T2, but prefer to declare untruthfully in an environment such as T3. In contrast, the endogenous audit probability p_i^H is higher than 0.5 if the subject i expects to be assessed as being less likely to be honest than the median among all other cheating subjects. For these individuals, the rather high endogenous audit probabilities result in a more compliant behavior in T3 than in T1 or T2. Individuals of this type are called 'weak liars'. Sufficiently weak liars, who may not declare truthfully in T1 or T2, may prefer to declare truthfully in an environment such as T3.

Note that the simultaneous presence of weak and strong liars causes a problem for identifying the treatment effect between T2 and T3. Weak liars can be expected to show higher compliance in T3 than in T2. Strong liars can be expected to be less compliant in T3 than they are in T2. The presence of two groups of equal size would then cause deviations among the strong and the weak liars that may lead to very similar aggregate behavior in T2 and T3. We addressed this problem as follows. We used two different parameter settings for the fine θ for all treatments. In half of all sessions the fine was low ($\theta = 100$), in the other half of the sessions the fine was high ($\theta = 300$). The low-fine treatments are suitable to identify the effect of weak liars, whereas the high-fine treatments are suitable to identify the effect of strong liars.

To see this, consider the indifference condition (2) for θ for individuals who are motivated by their monetary payoffs. It shows that the low (high) fine makes cheating a bet with a higher (lower) expected payoff than truthful compliance for high-endowment individuals. If individuals maximize their expected monetary payoff and have a high endowment ($x_i = 1000$), they should declare honestly ($y_i^H = h$) in case of a high fine ($T < \theta = 300$), but cheat ($y_i^H = l$) in case of a low fine ($T > \theta = 100$) for both treatments T1 and T2. The difference in punishment size should have a qualitatively similar effect in T3. In the treatments with a low fine, both average liars and strong liars have a monetary incentive to cheat. Only very weak liars may find it preferable to declare truthfully. Similarly, for a high fine, both average and weak liars have a monetary incentive to declare truthfully, but sufficiently strong liars may find it attractive to cheat. Hence, the low-fine setting ($T > \theta = 100$) is chosen to measure the existence of weak liars from the difference in compliance behavior between T2 and T3, and the high-fine setting ($T < \theta = 300$) is chosen to infer and measure the existence of strong liars from the difference in compliance behavior between T2 and T3.

In the experiment, we expect that choices are made not only on the basis of material rewards.

Other idiosyncratic factors are likely to play a role in an individual's compliance decision. We, therefore, do not expect these predictions to materialize sharply, but that there is some noise in the data.[20]

Hypotheses

Before turning to the comparison of the baseline treatment T1 with T3, we test for the role of pure face-to-face contact in the compliance decision. As touched above, the treatment T2 serves as a control treatment since it captures changes in compliance behavior that only arise because individuals are now required to make their declaration decision in a face-to-face situation, with the audit mechanism remaining the same as in T1. For individuals who are motivated by the monetary payoffs, the declaration mode (face-to-face in T2 or via a computer in T1) should not matter, as long as the audit probability is exogenous. We formulate this as an auxiliary hypothesis:

There is no treatment effect between T1 and T2 for subjects with a high endowment ($x_i = 1000$).[21]

The hypothesis is auxiliary to what we do, as it paves the way for our main research question. Given the behavioral literature on individuals' attitudes toward lying under a variety of conditions (Lundquist et al. 2009), a competing (behavioral) hypothesis suggests that the share of high-endowment subjects declaring honestly is higher in T2 than in T1 since, e.g., lying costs are higher in a face-to-face setting. A substantial change in compliance behavior between T1 and T2 would suggest that a change in compliance behavior from T1 and T3 is driven by a mixture of a face-to-face effect and belief formation about the endogenous audit probability. If that is the case, the effect of second-order beliefs is captured by the difference between T2 and T3. If, however, the change in compliance behavior between T1 and T2 is small, this implies that the difference in behavior from T1 to T3 can be mostly attributed to the effect of the endogenous audit probability. To sum up, behavioral aspects, such as an aversion against lying, materialize in the data in a comparison between T1 and T2, but a difference between T2 and T3 must be attributed to the role of beliefs about own audit probability.

For a comparison between T1 and T3, we take into consideration that individuals differ in exogenous characteristics. These characteristics can be aggregated in what we label a subject's

[20] This comparison also tests whether the standard theory results on the effectiveness of higher fines hold in our framework. But, this test is not central to our research question. Given the considerable evidence on earlier tax compliance experiments, we expected that the size of the fine matters.

[21] A more formal underpinning of the whole analysis, which considers the compliance decision of individuals with exogenous and endogenous audit probabilities as a Bayesian game, is presented in the appendix of an earlier version of this paper; cf. Konrad, Lohse and Qari (2012). This auxiliary hypothesis is derived in that appendix as Proposition 1.

look of being honest.[22] We do not consider neither what this look means precisely, whether this is an objective feature that induces the subject to be more honest in equilibrium than others, nor whether this quality exists only in the imagination of a subject. Suppose that subjects cannot alter their own look and that this look is the quasi automated basis for selecting subjects for an audit. If all subjects have a precise idea about whether they have an "honest look" or a "dishonest look" and have prior beliefs about the distribution of these looks in the subject pool, they form conjectures regarding whether subjects with a certain look report truthfully or underreport. This translates into a distribution of looks in the subset of individuals who actually underreport. By construction, the customs officer does not consistently solve for an equilibrium, but simply sorts subjects according to their looks. Half of the subjects in this subset are automatically subjected to an audit. This half is, by construction, the less-honest-looking half in the eyes of the customs officer.[23] For the numerical case with high fines ($x^H = 1000\,; T = 200; \theta = 300$), the critical level of p_i^H, for which i is indifferent about whether to report truthfully, is $p_i^H = 2/5$ by condition (2). Accordingly, if i, as a material-payoff-motivated subject, considers her endogenous audit probability to be below 40%, the equilibrium prediction is that she, as a particularly honest-looking subject, will underreport in treatment T3 with a high fine, whereas, for the same high fines, the prediction for T1 and T2 is that no subject should underreport for this parameter range. This yields our first main testable hypothesis.

Hypothesis A (Existence of strong liars): *In a high-fine setting ($T<\theta = 300$), the share of individuals with a high endowment ($x_i = x^H$) who declare honestly in T3, is smaller than in T1 or T2.*

For the numerical case in the experimental setting of T3 with low fines ($x^H = 1000\,; T = 200; \theta = 100$), the critical level of p_i^H, for which the subject is indifferent about reporting truthfully or not, is $p_i^H = 2/3$ by condition (2). Accordingly, the equilibrium prediction for material-payoff-motivated subjects is that some (particularly dishonest-looking) subjects will report truthfully in treatment T3 with low fines, whereas, for the same low fines, the prediction for T1 and T2 was that no material-payoff-motivated subject should report truthfully for this level of the fine. This yields our second main testable hypothesis.

Hypothesis B (Existence of weak liars): *In a low-fine setting ($T >\theta = 100$), the share of individuals with a high endowment ($x_i = x^H$), who declare honestly in T3, is larger than in T1 or T2.*

[22] The questions of whether an individual is aware of her look and if this self-assessment about her look is consistent with the perceptions by others is studied by Konrad, Lohse and Qari (2014).

[23] A more formal analysis is presented in the appendix of Konrad et al. (2012), and the two hypotheses A and B follow directly from the characterization of the Bayesian Nash equilibrium in Proposition 2 in that appendix.

Evidence in favor of hypothesis A or B, at least one of them, suggests that the endogenous audit probability has an effect on the compliance behavior.

3 Results

In this section we analyze subjects' aggregate behavior across treatments using different econometric models. We first briefly describe the characteristics of our sample. Then we explain the associated empirical strategy. We then estimate treatment effects according to the between-subjects design outlined in the previous section.

Sample characteristics and empirical strategy

As described in section 2, we carried out 4 sessions with the baseline treatment (T1), 4 with T2 and 4 with T3. In one half of the sessions we have $T > \theta$, such that dishonest behavior is induced, while, in the other half, incentives to report a high endowment honestly are induced ($T < \theta$). T1 is played ten rounds, while the number of rounds equals four in T2 and T3. Since there are 20 participants in each session, there are in total 80 subjects and 800 observations from T1, while there are 160 subjects with 640 corresponding observations from T2 and T3. Note that subjects' endowments were randomly assigned –with replacement– in each round. Hence, the number of low- and high-endowment observations, respectively, is not fixed ex-ante.[24]

Table 1 provides a first summary of the sample characteristics. For the data analysis, the reporting variable y_{it} is coded as follows: y_{it} is equal to 0 if subject i in period t reports a low endowment and equal to 1 if subject i reports a high endowment. The upper panel tabulates the number of low-endowment reports (y_{it}=0) by treatment and true endowment x_{it}, and the lower panel tabulates the high-endowment reports. As discussed earlier, subjects are never

Table 1: Reporting behavior by true endowment

Declaration y_{it}	Treatment	True endowment x_{it}	
		$x^L = 400$	$x^H = 1000$
0 ($y_{it} = l$)	T1	158	341
0 ($y_{it} = l$)	T2	64	106
0 ($y_{it} = l$)	T3	60	92
1 ($y_{it} = h$)	T1	3	298
1 ($y_{it} = h$)	T2	0	150
1 ($y_{it} = h$)	T3	0	168

expected to report a high endowment when their true endowment is low. As shown in the

[24] Due to the large number of observations, the fraction of low endowment observations is very close to 20% overall and as well in each treatment. The largest deviation from 20% occurs in T3, where the fraction of low endowment observations equals 0.1875.

table, there are only 3 observations that fit into this category. Hence, this provides a first indication that the subjects understood the rules of the game correctly. Recall that the low-endowment observations are only used to generate a meaningful experimental setup. For the evaluation of our main hypotheses, we do not need these observations. We therefore analyze only high endowment observations (where $x_{it} = x^H = 1000$) in the following.

We ask whether the compliance rate varies systematically across treatments. Our basic regression equation reads

$$y_{it} = \mathbf{\Xi'_{it}}\boldsymbol{\beta} + u_i + \epsilon_{it}, \tag{3}$$

where the binary variable y_{it} is equal to one, if subject i, who has a high endowment in round t, truthfully reports this endowment. The main explanatory variables, collected in $\mathbf{\Xi'_{it}}$, are a series of dummy variables indicating the treatment in which subject i participated. The subject-specific error term u_i controls for the repeated measurement of each subject.

Note that the sample is unbalanced in two respects. First, subjects in T1 play 10 rounds, while subjects in T2 and T3 play 4 rounds.[25] Second, the number of high-endowment observations differs across subjects. As the probability of having a high endowment equals 80%, most subjects in T2 and T3 obtained a high endowment in at least three rounds. This implies that most subjects are observed either three or four times in those treatments and few subjects are observed once or twice. We therefore use the well-established parametric approach[26] to handle such sampling conditions and employ a nonlinear (logistic) mixed effects model[27] to fit equation (3), i.e. the model predicts the probability of reporting truthful as follows: $P(Y_{it} = 1|u_i, \mathbf{\Xi'_{it}}) = \Lambda\left(\mathbf{\Xi'_{it}}\boldsymbol{\beta} + u_i\right)$ where $\Lambda(\cdot)$ denotes the inverse logit function.

Treatment effects

Table 2 and Figure 2 present the aggregate compliance rate by treatment and penalty. Starting the discussion with the low penalty setup ($T > \theta$), we find that the compliance rates in the first two treatments are very similar. The fraction of truthful reports is roughly 28% in T1 and

[25] We carry out robustness checks by using only the first or the last 4 rounds from T1 for the estimation of treatment effects (see Section 4 for details). This does not affect our main evidence.

[26] While rank-based methods are well developed and known to be very efficient for the case of independent balanced data, there is no well-established procedure for the case of unbalanced data involving clusters due to the repeated measurement of the same subjects. A common remedy is to reduce the dataset by calculating the average response for each subject. Unlike the raw data, the averages are independent and do not follow a dichotomous distribution. However, since the number of observations entering the respective averages varies across the sample, using a Wilcoxon rank-sum test is likely to generate test statistics of improper size (e.g. Datta and Satten 2005).

[27] The term "mixed effects model" refers to the fact that both fixed and random effects are estimated. See, for example, Agresti (2003), Wooldridge (2006) or Cameron and Miller (2009). Since we fit only one parameter for the subject-specific intercepts, the model is equivalent to what the economics literature calls a "random effects logit model".

Figure 2: Reporting by treatment and penalty

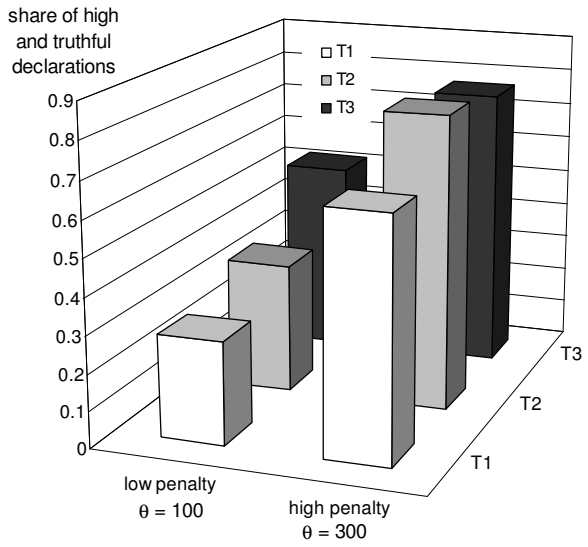


Table 2: Share of high-endowment reports (among high-endowment cases)

	Penalty		
Treatment	Low ($T > \theta$)	High ($T < \theta$)	N
1	.28	.65	639
2	.35	.80	256
3	.52	.77	260
N	575	580	1150

35% in T2. Hence, the baseline compliance rate is "too high" if subjects consider only expected monetary payoffs. As mentioned above, we do not expect hypotheses A and B to materialize sharply. In fact, this finding resembles results from previous tax compliance experiments (Alm and Jacobson 2007). Moving to the treatment with endogenous audit probabilities, we observe that the compliance rate is considerably higher (close to 52%) than in T1 and T2. This suggests that the influence of mere face-to-face contact is much smaller than the effect of the endogenous audit probability. This evidence is in line with Hypothesis B: For $\theta = 100$, the compliance rate in T3 is substantially larger than in T1 and T2, i.e. there is evidence of an increase in compliance, driven by the fact that the audit probability is endogenous in T3.

Turning to the high-penalty setup ($T < \theta$), a first finding is that the baseline compliance rate in T1 is equal to 65%. This rate is too small compared to the expected rate of 100% if subjects consider only expected monetary payoffs. Once again, this resembles the recurring finding in tax compliance experiments that changes in compliance behavior in response to higher fines are often too small and are usually smaller than changes in response to higher audit probabilities, even if expected payoffs of the changes are the same (Alm and Jacobson 2007, Alm 2012). Compared to this baseline compliance rate, the share of honest reports is higher in T2 than in T1, although the audit probability is constant across these two treatments. Furthermore, unlike in the low-penalty case, the compliance rate in T2 is seemingly the same as in T3. In summary, the descriptive evidence supports Hypothesis B (existence of weak liars) while there is no support for Hypothesis A (existence of strong liars).

We now move to the econometric evidence and fit the logistic version of equation (3) by maximum likelihood.[28] Tables 3 and 4 summarize the results for the low- and high-penalty setup, respectively. The first column presents the results for the full model including treatment dummies and individual random effects, while the second column estimates a reduced model omitting the treatment dummies. We discuss the reduced model in more detail below. Overall, the predictions obtained from the logit model confirm the descriptive evidence. In the low-penalty case (Table 3), the predicted fraction of truthful reports in T1 is equal to 29%. Compared to T1, the predicted compliance rate in the intermediate treatment T2 is five percentage points higher, generating a compliance rate of about 36%. However, the standard error (0.707) of the T2-coefficent (0.426) reveals that this difference is not statistically significant. The predicted compliance rate in T3 is roughly 80% higher than in T1, yielding a total compliance rate of 51%. The intermediate treatment T2 captures the effect of mere face-to-face contact while holding the monetary incentives constant, and, hence, the pure face-to-face effect seems to be rather small for the low-penalty setup. In turn, this suggests that the increase in

[28] We use the R environment (R Development Core Team 2009) and in particular the lme4 package (Bates and Maechler 2009) to fit the model.

tax compliance in T3 is driven by the endogenous audit probabilities. To summarize our results so far, the evidence in the low-penalty setup is in line with Hypothesis B and the auxiliary hypothesis of no treatment effect between T1 and T2.

Table 3: Logistic mixed effects model for $\theta = 100$ (low penalty)

	(1)	(2)
(Intercept)	-1.753^{***}	-1.080^{***}
	(0.476)	(0.303)
Treatment2	0.426	
	(0.707)	
Treatment3	1.872^{**}	
	(0.698)	
σ_u^2	7.325	8.167
Log-likelihood	−290.141	−293.520
AIC	588.281	591.040
N	575	575

Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The table presents the results from a logistic mixed effects model where the probability of reporting truthfully $P(Y_{it} = 1)$ is predicted by a set of treatment fixed effects (dummy variables) and a random intercept for each subject. The omitted reference category in column (1) is Treatment 1. Column (2) presents the results from a more parsimonious model where y_{it} depends only on the subject-specific random effect and an intercept.

The regression results for the high-penalty setup (Table 4) also resemble the descriptive evidence. The compliance rate in T1 is roughly 65%, and it is between 15 and 18 percentage points higher in T2 and T3, yielding a total compliance rate of about 80-83%.

We can assess the overall explanatory power of the treatment dummies by comparing the model to a model including only an intercept and the random effects. Formally, the equation of the more parsimonious model reads $y_{it} = b + u_i + \epsilon_{it}$. The model is nested in the previous model allowing a likelihood-based comparison of both models. For both the low and high penalty setup, the comparison of the respective Akaike information criteria (or the associated Likelihood ratios) suggests that the model including treatment dummies is preferred. However, as we discuss in the following section, the estimates for the high-penalty setup are not robust and are insignifcant in most of the following specifications.

Overall, while we find no evidence for the presence of strong liars in the high-penalty case, there is clearly a large fraction of weak liars in the low-penalty case. Moreover, the 80% increase of compliance in the low-penalty case can be attributed to the formation of subjective audit probabilities rather than to a pure face-to-face effect. Indeed, the coefficient of the control treatment (T2) that captures mere face-to-face effects is imprecisely estimated; the compliance

Table 4: Logistic mixed effects model for $\theta = 300$ (high penalty)

	(1)	(2)
(Intercept)	1.477*	2.603***
	(0.590)	(0.368)
Treatment2	2.207*	
	(0.966)	
Treatment3	1.802	
	(0.933)	
σ_u^2	11.726	10.860
Log-likelihood	−260.545	−263.843
AIC	529.090	531.686
N	580	580

Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The table presents the results from a logistic mixed effects model where the probability of reporting truthfully $P(Y_{it} = 1)$ is predicted by a set of treatment fixed effects (dummy variables) and a random intercept for each subject. The omitted reference category in column (1) is Treatment 1. Column (2) presents the results from a more parsimonious model where y_{it} depends only on the subject-specific random effect and an intercept.

rate for the computerized treatment (T1) and the control treatment (T2) are statistically not distinguishable. The following section presents a series of checks to inquire as to the robustness of these results. In particular, we include controls for individual risk attitudes, time effects, and further controls. The main conclusions of this section are not affected by these robustness checks.

4 Robustness checks

In this section we inquire as to the robustness of the estimated treatment effects. As before, all results in this section rely on logistic unobserved effects models that take into account both the binary response variable and the repeated measurement of the same individuals.

Variation over time and risk attitudes

Figures S1 and S2 in the online appendix depict the compliance rate in each round averaging over subjects for the low- and high-penalty setup respectively. Especially for the former case, there is some evidence that the compliance rate decreases in later periods. To facilitate a comparison across the treatments, column (2) of Tables S1 and S2 (in the online appendix) respectively enters a dummy variable indicating the second half (*laterRounds*) of the game

into the regression. We also introduce the individual measure of risk aversion, measured by carrying out a game in the style of Holt and Laury (2002), to the model.[29]

The estimated treatment effects in the low-penalty setup (Table S1 in the online appendix) are robust with regard to the inclusion of these additional controls. Moreover, the coefficient for the risk measure is quite small, and the estimated time effect obtains a large standard error. Hence, both additional controls have no predictive power in the low-penalty case.

The risk coefficient in the high-penalty setup (Table S2 in the online appendix) obtains a value of 0.583. As expected from the figures, the estimated time coefficient is larger compared to the low-penalty case. Moreover, entering the risk measure drives down the treatment coefficients which thereby become insignificant. Hence, the estimates for high penalty setup are not robust to the inclusion of these controls. Column (3) of the two tables considers a linear time trend instead of the dummy variable. Note that the coefficient in this specification (*round*) is driven mainly by T1 since high values for this variable only apply in the first treatment. This approach lowers the estimated treatment coefficients in both the low- and high-penalty setups. However, the qualitative evidence remains the same.

Finally, we inquire the robustness of the results with respect to the different number of rounds in T1 and T2/T3. In the first set of models, we use the first four (out of 10) rounds from T1 to estimate treatments effects, while we use the last four rounds in the second set of models. Table S3 compiles the results. Columns (1) and (2) present the results for the low-penalty setup. Both models corroborate our main results. There is only a small (and insignificant) difference between T1 and T2, while the compliance rate in T3 is much larger. Columns (3) and (4) present the results for high-penalty case. As before, the two regressions indicate that there is not much variation across treatments in the high-penalty setup. For example, the model using the first four rounds (column 3) once again indicates that the compliance rate in T2 and T3 are not significantly different from the compliance rate in T1.

Gender and age effects

Columns (4) and (5) of Tables S1 and S2 introduce a gender dummy and an age effect, respectively, as additional control variables. The gender coefficient in the low-penalty setup (Table S1) is slightly larger (around 0.788) than the coefficient for the risk measure. However, the standard errors indicate that the precision of these estimates is rather poor. The age coefficient (column 5) is also noisy and small. In the high-penalty setup (Table S2) the gender and the age coefficients are slightly smaller and also imprecise. More importantly, the estimates for the

[29] The fraction of risk-taking subjects is roughly 10% (where "risk-taking" means that values for the risk measure between 1 and 4 are obtained). The fraction of subjects preferring medium risk is 48% (risk measure 5-7); the fraction of risk-averse (risk measure 8-10) subjects is 42%.

treatment effects are in line with the results so far: In the high-penalty setup, the treatment dummies are not significant, while in the low-penalty setup the estimate for the T3-coefficient is large and significant.

Using T2 as the reference treatment

Note that the econometric models presented so far employed T1 as the reference treatment. However, one might interpret T1 and T3 as variations of T2, which therefore could be alternatively considered as the reference treatment. While the difference between T2 and T1 is readily available in Tables S1 and S2 respectively, there is no direct estimate for the difference between T3 and T2 yet. We therefore run a logit model where T2 is the omitted base category.

Table S4 in the online appendix summarizes the results. The first column shows the coefficients for $\theta = 100$. The standard error for the difference between T2 and T1 is fairly large. This suggests that the small difference in the average compliance rate between T2 and T1 (five percentage points) might be attributed to sampling error. By contrast, the coefficient for the difference between T3 and T2 (0.69) is precisely estimated and the associated $z-$ratio indicates statistical significance on the 5%-level. Column (2) presents the coefficients for $\theta = 300$. In line with previous results, the estimates suggest that, in this high penalty setup, aggregate compliance rates in T2 and T3 are similar and higher compared to T1.

Robustness checks: summary

Summarizing, all models corroborate the descriptive evidence: For $\theta = 100$ aggregate tax compliance in T3 is considerably higher than in T1. Tax compliance is similar in T1 and T2. Hence, there is a compliance-increasing effect, driven by the introduction of endogenous audit probabilities. For $\theta = 300$ the treatment dummies do not explain much variation and are not significant in most specifications. Recall that all results are obtained from models including random effects to control for heterogeneity of subjects.

5 Conclusion

We analyze the role of endogenous audit probabilities in a compliance framework, taking as an example the situation at customs. Our experimental results reveal a major asymmetry: On the one hand, a considerable number of subjects assess their audit probability as rather high. We refer to this type of subjects as weak liars. On the other hand, there is no evidence of strong liars, i.e. of subjects who behave as if they consider their audit probability to be rather low.

In one half of all experimental sessions, the subjects report in compliance frameworks in which low fines are a monetary incentive to underreport. In one treatment (T3), customs officers make assessments on the basis of personal communication. In this treatment the subjects' appearance and performance influence who among them receives an audit. Aggregate compliance behavior is substantially higher (about 80%) in this treatment, as compared to a treatment in which declaration occurs via a computer with a strictly random audit (T1). Using a further control treatment (T2), we can also distinguish between the role of pure face-to-face contact and the role of customs officers' assessments and the formation of endogenous audit probabilities. Our findings are in line with the hypothesis that that higher compliance is in fact driven by endogenous probabilities.

In the other half of all sessions, we provide subjects with monetary incentives to report truthfully: high fines. In such an environment, only those subjects would have incentives to underreport, who believe that they can successfully fool the customs officer. In this experiment we find no evidence for such behavior as the estimates for the high-penalty setup are not robust and are insignificant in most of the specifications.

In interviews with customs officers from three international German airports, about 50% indicated that they consider personal interviews to be the most effective strategy to increase customs compliance. Our experimental evidence is in line with this perception of the existence of weak liars. In contrast, customs officers also stated that there are strong liars, for whom we can not find evidence.

Our results - hardly any difference in compliance behavior between T1 and T2, but between those treatments and T3 - indicate that the value of a face as observed by Eckel and Petrie (2011) does not result from a dislike of anonymity per se since this would imply behavioral changes between T1 and T2. In fact, the face value might be traced back to a preference for unveiling the counterpart's face when interacting economically as it happens in T3.

Our findings about the role of the endogenous audit probability have policy relevance not only for the institutional framework of customs declarations. The results are also supportive of the institutional set-up of tax declarations and other compliance frameworks more generally: Truthful compliance may possibly be increased, if the declarations or reports are made in person and if the audit probability for subjects is influenced by their appearance and performance.

References

Agresti, A. (2003). *Categorical Data Analysis*, John Wiley & Sons, Inc.

Allingham, M. G. and Sandmo, A. (1972). Income tax evasion: A theoretical analysis, *Journal of Public Economics* 1(3-4): 323–338.

Alm, J. (2010). Testing behavioral public economics theories in the laboratory, *National Tax Journal* 63(4): 635–658.

Alm, J. (2012). Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies, *International Tax and Public Finance* 19(1): 54–77.

Alm, J., Cherry, T., Jones, M. and McKee, M. (2010). Taxpayer information assistance services and tax compliance behavior, *Journal of Economic Psychology* 31(4): 577–586.

Alm, J. and Jacobson, S. (2007). Using laboratory experiments in public economics, *National Tax Journal* 60(1): 129.

Andreoni, J., Erard, B. and Feinstein, J. (1998). Tax compliance, *Journal of Economic Literature* 36(2): 818–860.

Andreoni, J. and Petrie, R. (2008). Beauty, gender and stereotypes: Evidence from laboratory experiments, *Journal of Economic Psychology* 29(1): 73–93.

Bates, D. and Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-32.

Becker, G. S. (1968). Crime and punishment: An economic approach, *Journal of Political Economy* 76(2): 169.

Boadway, R., Marceau, N. and Mongrain, S. (2007). Redistributive taxation under ethical behaviour, *The Scandinavian Journal of Economics* 109(3): 505–529.

Camerer, C. F., Ho, T. H. and Chong, J.-K. (2004). A cognitive hierarchy model of games, *The Quarterly Journal of Economics* 119(3): 861–898.

Cameron, A. C. and Miller, D. L. (2009). Robust inference with clustered data, Working Papers 107, University of California, Davis, Department of Economics.

Castillo, M. and Petrie, R. (2010). Discrimination in the lab: Does information trump appearance?, *Games and Economic Behavior* 68(1): 50–59.

Chander, P. and Wilde, L. L. (1998). A general characterization of optimal income tax enforcement, *Review of Economic Studies* 65(1): 165–83.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership, *Econometrica* 74(6): 1579–1601.

Charness, G., Gneezy, U. and Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design, *Journal of Economic Behavior & Organization* 81(1): 1–8.

Coricelli, G., Joffily, M., Montmarquette, C. and Villeval, M. (2010). Cheating, emotions, and rationality: an experiment on tax evasion, *Experimental Economics* 13(2): 226–247.

Costa-Gomes, M. A. and Weizsäcker, G. (2008). Stated beliefs and play in normal-form games, *Review of Economic Studies* 75(3): 729–762.

Datta, S. and Satten, G. A. (2005). Rank-sum tests for clustered data, *Journal of the American Statistical Association* 100(471): 908–915.

Eckel, C. C. and Petrie, R. (2011). Face value, *American Economic Review* 101(4): 1497–1513.

Feldman, N. E. and Slemrod, J. (2007). Estimating tax noncompliance with evidence from unaudited tax returns, *Economic Journal* 117(518): 327–352.

Frey, B. (1997). A constitution for knaves crowds out civic virtues, *The Economic Journal* pp. 1043–1053.

Frey, B. S. and Torgler, B. (2007). Tax morale and conditional cooperation, *Journal of Comparative Economics* 35(1): 136–159.

Greiner, B. (2004). An online recruitment system for economic experiments, MPRA Paper 13513, University Library of Munich, Germany.

Habyarimana, J., Humphreys, M., Posner, D. and Weinstein, J. (2007). Why does ethnic diversity undermine public goods provision?, *American Political Science Review* 101(4): 709.

Hartner, M., Rechberger, S., Kirchler, E. and Schabmann, A. (2008). Procedural fairness and tax compliance, *Economic Analysis and Policy (EAP)* 38(1): 137–152.

Holm, H. J. and Kawagoe, T. (2010). Face-to-face lying - an experimental study in Sweden and Japan, *Journal of Economic Psychology* 31(3): 310–321.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects, *American Economic Review* 92(5): 1644–1655.

Kleven, H., Knudsen, M., Kreiner, C., Pedersen, S. and Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark, *Econometrica* 79(3): 651–692.

Konrad, K. A., Lohse, T. and Qari, S. (2012). Customs compliance and the power of imagination, CESifo Working Paper Series 3702, CESifo Group Munich.

Konrad, K. A., Lohse, T. and Qari, S. (2014). Deception choice and self-selection - the importance of being earnest, *Journal of Economic Behavior & Organization* 107A: 25–39.

Konrad, K. A. and Qari, S. (2012). The last refuge of a scoundrel? patriotism and tax compliance, *Economica* 79(315): 516–533.

Lohse, T. and Qari, S. (2014). Gender differences in deception behaviour–the role of the counterpart, *Applied Economics Letters* 21(10): 702–705.

Lundquist, T., Ellingsen, T., Gribbe, E. and Johannesson, M. (2009). The aversion to lying, *Journal of Economic Behavior & Organization* 70(1-2): 81–92.

Manski, C. and Neri, C. (2011). First- and second-order subjective expectations in strategic decision-making: Experimental evidence, mimeo.

Mobius, M. M. and Rosenblat, T. S. (2006). Why beauty matters, *American Economic Review* 96(1): 222–235.

R Development Core Team (2009). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem, *Econometrica* 68(5): 1281–1292.

Reinganum, J. F. and Wilde, L. L. (1985). Income tax compliance in a principal-agent framework, *Journal of Public Economics* 26(1): 1–18.

Reinganum, J. F. and Wilde, L. L. (1986). Equilibrium verification and reporting policies in a model of tax compliance, *International Economic Review* 27(3): 739–760.

Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion, *Journal of Economic Perspectives* 21(1): 25–48.

Solnick, S. J. and Schweitzer, M. E. (1999). The influence of physical attractiveness and gender on ultimatum game decisions, *Organizational Behavior and Human Decision Processes* 79(3): 199–215.

Thursby, M., Jensen, R. and Thursby, J. (1991). Smuggling, camouflaging, and market structure, *The Quarterly Journal of Economics* 106(3): 789–814.

Torgler, B. (2006). The importance of faith: Tax morale and religiosity, *Journal of Economic Behavior & Organization* 61(1): 81–109.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations, *Econometrica* 76(6): 1467–1480.

Weizsäcker, G. (2003). Ignoring the rationality of others: Evidence from experimental normal-form games, *Games and Economic Behavior* 44(1): 145–171.

Wilson, R. K. and Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game, *Political Research Quarterly* 59(2): 189–202.

Wooldridge, J. M. (2006). Cluster-sample methods in applied econometrics: An extended analysis, Technical report.

Yaniv, G. (2010). The red-green channel dilemma: Customs declaration and optimal inspection policy, *Review of International Economics* 18(3): 482–492.

Yitzhaki, S. (1974). Income tax evasion: A theoretical analysis, *Journal of Public Economics* 3(2): 201–202.

Supplementary Online Appendix

Table S1: Logistic mixed effects models (low penalty)

	(1)	(2)	(3)	(4)	(5)
(Intercept)	-1.753^{***}	-1.592^{**}	-1.085	-2.102^{**}	-2.118^{*}
	(0.476)	(0.494)	(0.558)	(0.640)	(1.035)
Treatment2	0.426	0.383	0.026	0.374	0.376
	(0.707)	(0.720)	(0.736)	(0.724)	(0.728)
Treatment3	1.872^{**}	1.890^{**}	1.532^{*}	1.855^{**}	1.856^{**}
	(0.698)	(0.704)	(0.719)	(0.703)	(0.704)
RiskMeasure		0.061	0.058	0.064	0.064
		(0.156)	(0.156)	(0.156)	(0.157)
laterRounds		-0.336		-0.338	-0.338
		(0.243)		(0.244)	(0.244)
round			-0.126^{*}		
			(0.055)		
Female				0.788	0.789
				(0.622)	(0.622)
Age					0.002
					(0.103)
σ_u^2	7.325	7.478	7.502	7.469	7.468
Log-likelihood	-290.141	-289.185	-287.550	-288.458	-288.458
AIC	588.281	590.371	587.100	590.916	592.916
N	575	575	575	575	575

Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The table presents the results from a logistic mixed effects model where the probability of reporting truthfully $P(Y_{it} = 1)$ is predicted by a set of treatment fixed effects (dummy variables) and a random intercept for each subject. The omitted reference category is Treatment 1.

Table S2: Logistic mixed effects models (high penalty)

	(1)	(2)	(3)	(4)	(5)
(Intercept)	1.477*	1.746**	2.152***	2.176**	1.392
	(0.590)	(0.578)	(0.643)	(0.779)	(1.214)
Treatment2	2.207*	1.591	1.266	1.531	1.565
	(0.966)	(0.930)	(0.943)	(0.941)	(0.934)
Treatment3	1.802	1.556	1.241	1.516	1.566
	(0.933)	(0.883)	(0.900)	(0.894)	(0.892)
RiskMeasure		0.583**	0.586**	0.614**	0.604**
		(0.214)	(0.214)	(0.219)	(0.217)
laterRounds		−0.329		−0.328	−0.326
		(0.269)		(0.269)	(0.269)
round			−0.104		
			(0.056)		
Female				−0.639	−0.631
				(0.788)	(0.780)
Age					0.092
					(0.113)
σ_u^2	11.726	10.271	10.239	10.426	10.208
Log-likelihood	−260.545	−255.781	−254.843	−255.421	−255.065
AIC	529.090	523.562	521.686	524.841	526.129
N	580	580	580	580	580

Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The table presents the results from a logistic mixed effects model where the probability of reporting truthfully $P(Y_{it} = 1)$ is predicted by a set of treatment fixed effects (dummy variables) and a random intercept for each subject. The omitted reference category is Treatment 1.

Figure S1: Reporting by treatment and rounds (low penalty)

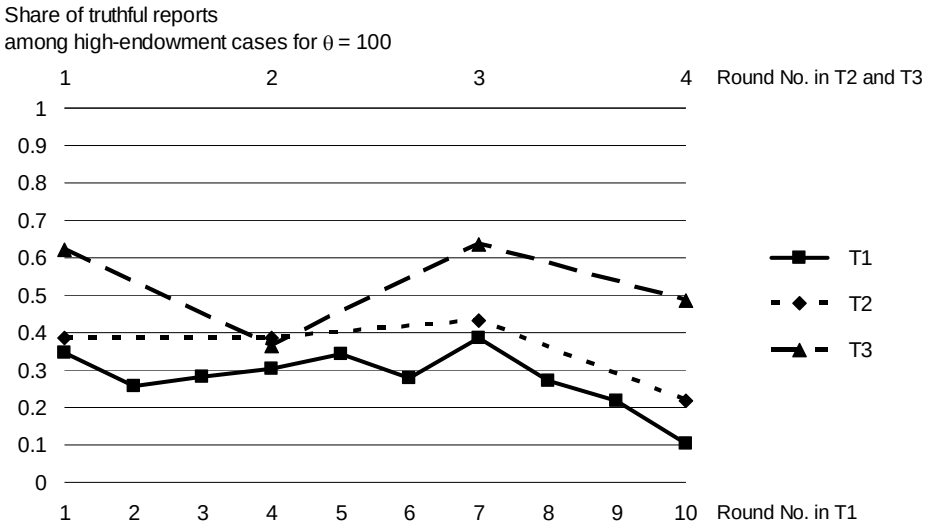


Figure 2: Reporting behavior among high-endowment cases for Treatments T1 (black), T2 (dotted) and T3 (dashed) for θ = 100. Note that the fully computerized treatment T1 had 10 rounds whereas, treatments T2 and T3, with personal interviews, had only four rounds, but each of the rounds in T2 and T3 took more time and probably had more salience, meaning that the learning effects from the first to the last round in the different treatments may, nevertheless, be comparable.

Table S3: Logistic mixed effects models (first and last four rounds from T1)

	(1) $\theta = 100$ (first)	(2) $\theta = 100$ (last)	(3) $\theta = 300$ (first)	(4) $\theta = 300$ (last)
(Intercept)	-1.767^{**}	-2.384^{***}	2.827^{**}	0.933
	(0.584)	(0.617)	(0.931)	(0.526)
Treatment2	0.427	1.037	1.520	1.911^{*}
	(0.825)	(0.851)	(1.433)	(0.809)
Treatment3	1.892^{*}	2.509^{**}	1.142	1.571^{*}
	(0.817)	(0.842)	(1.390)	(0.790)
σ_u^2	9.481	9.600	25.040	8.010
Log-likelihood	−195.121	−190.405	−171.868	−181.292
AIC	398.242	388.810	351.735	370.584
N	379	379	393	387

Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The table presents the results from a logistic mixed effects model where the probability of reporting truthfully $P(Y_{it} = 1)$ is predicted by a set of treatment fixed effects (dummy variables) and a random intercept for each subject. The omitted reference category is Treatment 1. Columns (1) and (3) use the first four (out of ten) rounds from Treatment 1, while columns (2) and (4) use the last four rounds from Treatment 1.

Figure S2: Reporting by treatment and rounds (high penalty)

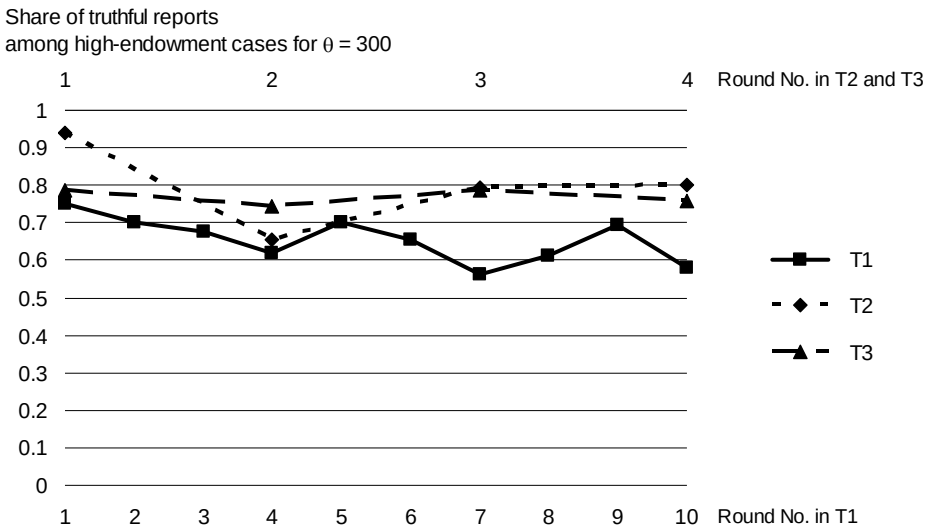


Figure 3: Reporting behavior among high-endowment cases for Treatments T1 (black), T2 (dotted) and T3 (dashed) for θ = 300. Note that the fully computerized treatment T1 had 10 rounds, whereas treatments T2 and T3, with personal interviews, had only four rounds, but each of the rounds in T2 and T3 took more time and probably had more salience, meaning that the learning effects from the first to the last round in the different treatments may, nevertheless, be comparable.

Table S4: Logistic mixed effects model (T2 is the omitted baseline category)

	(1) $\theta = 100$	(2) $\theta = 300$
(Intercept)	-1.327^{*}	3.684^{***}
	(0.523)	(0.766)
T1-T2	-0.426	-2.207^{*}
	(0.707)	(0.966)
T3-T2	1.446^{*}	-0.405
	(0.730)	(1.053)
σ_u^2	7.325	11.726
Log-likelihood	-290.141	-260.545
AIC	588.281	529.090
N	575	580

Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The table presents the results from a logistic mixed effects model where the probability of reporting truthfully $P(Y_{it} = 1)$ is predicted by a set of treatment fixed effects (dummy variables) and a random intercept for each subject. The omitted reference category is Treatment 2.