ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

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

Working Paper

Customs compliance and the power of imagination

WZB Discussion Paper, No. SP II 2011-108

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Konrad, Kai A.; Lohse, Tim; Qari, Salmai (2011) : Customs compliance and the power of imagination, WZB Discussion Paper, No. SP II 2011-108, Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin

This Version is available at:
https://hdl.handle.net/10419/54588

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Kai A. Konrad
Tim Lohse
Salmai Qari

Customs Compliance and the Power of Imagination

Discussion Paper

SP II 2011–108

December 2011

Social Science Research Center Berlin (WZB)

Research Area
Markets and Politics

Research Professorship & Project
The Future of Fiscal Federalism

Wissenschaftszentrum Berlin für Sozialforschung gGmbH
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Copyright remains with the author(s).

Discussion papers of the WZB serve to disseminate the research results of work in progress prior to publication to encourage the exchange of ideas and academic debate. Inclusion of a paper in the discussion paper series does not constitute publication and should not limit publication in any other venue. The discussion papers published by the WZB represent the views of the respective author(s) and not of the institute as a whole.

Affiliation of the authors other than WZB:

Kai A. Konrad
WZB and Max Planck Institute for Tax Law and Public Finance
MPI for Tax Law and Public Finance, Marstallplatz 1, 80539 Munich, Germany

Tim Lohse
Berlin School of Economics and Law and WZB
Berlin School of Economics and Law, Alt-Friedrichsfelde 60, 10315 Berlin, Germany

Salmai Qari
WZB and Max Planck Institute for Tax Law and Public Finance
MPI for Tax Law and Public Finance, Marstallplatz 1, 80539 Munich, Germany

Abstract

Customs Compliance and the Power of Imagination

Kai A. Konrad, Tim Lohse and Salmai Qari*

This paper studies the role of beliefs about own performance or appearance for compliance at the customs. In an experiment in which underreporting has a higher expected payoff than truthful reporting we find: a large share, about 15–20 percent of the subjects, is more compliant if they have reason to imagine that their performance influences their subjective audit probability. In contrast, we do not find evidence for individuals who believe that by their personal performance they can reduce the subjective probability for an audit. Our results suggest that the power of imagination, i.e. the role of second-order beliefs in the process of customs declarations is important and may potentially be used to improve customs and tax compliance.

Keywords: Customs, tax compliance, audit probability, second–order beliefs

JEL classification: H26, H31, C91

* For providing laboratory resources we kindly thank MELESSA of the University of Munich. We thank Hans Mueller for developing and programming the web-based environment and we thank Nina Bonge for creating the graphs. We thank Sophia Baur, Bernhard Enzi, Sabrina Korreck, Tobias Mirlach and Jennifer Zeiser for excellent research assistance, and we thank Elias Brumm and Thomas Daske for helping us to test the web environment. We also thank Ralph Bayer, Nadja Dwenger, Werner Güth, Changxia Ke, Florian Morath, Fangfang Tan, and seminar audiences at the Max Planck Institute for Tax Law and Public Finance and the University of Munich for helpful comments. We are very grateful to customs officers from the airports of Berlin-Tegel, Berlin-Schönefeld and Munich for giving us valuable insights into their work. The usual caveat applies.

1. Introduction

In this paper we investigate compliance decisions at the customs. If customs officers watch and decide whom to inspect, this may not leave travelers unaffected, particularly if they carry items which they are supposed to declare. We are interested in the role of travelers' imagination, i.e. their subjective beliefs about their own ability to influence customs officers. If the customs officer can choose whom to audit on the basis of the experience of personal interviews, travelers may imagine how their own physical characteristics (height, eye-color, ethnic background), their appearance (voice, looks, dress) and their eloquence will affect the customs officer's audit decision. Some individuals may feel that they have a low ability for deception. Other travelers may think that they have strong abilities to deceive customs officers, compared to the average ability for deception within the overall group of travelers. If both types of beliefs exist in a population of travelers, they work in opposite directions, and in the aggregate, this may cloud the true effect of such second-order beliefs. Our experimental design allows us to measure these effects separately.

Before designing the laboratory experiment, we conducted 17 interviews with customs officers at three different German international airports.[1] This served two purposes. First, it led to a better understanding of the existing framework at the customs. Second, we were interested in the self-perceptions of customs officers, particularly regarding their beliefs about the effectiveness of their audit policy. Key findings from the interviews, which were conducted 2009 and 2010, that are relevant for our research question are as follows. (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 about people getting nervous or turning red when just being watched by an officer. In case a traveler is chosen to be inspected, about half of the officers stated that they can tell right away whether or not the chosen traveler tried to smuggle some goods. (2) The perception was prevalent that the mere presence of customs officers induces a more honest behavior. Two thirds of the officers reported that their physical presence increased the share of travelers who chose to declare something. This suggests to design the experiment in such a way that we can separate the effects of second-order beliefs from the effect of a pure face-to-face contact. (3) We asked officers to rank three alternative institutional frameworks according to which framework is most effective to induce honest behavior. These included the voluntary self-selection of travelers into an exit for travelers who have "nothing to declare" and an exit for travelers who have to make a declaration, a written and signed declaration form, and verbal face-to-face customs declarations. About half of the officers interviewed considered

[1] The customs officers were encouraged by the customs administration to volunteer to 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.

verbal face-to-face declarations as most effective. (4) We find that customs officers have a large degree of freedom in their choices about whom to audit and develop their own heuristics on their job. This is important as it suggests that second-order beliefs have a legitimate place in real compliance contexts; travelers may rightly believe that their characteristics and behavior may affect their individual probability of being audited.

Our main findings can be briefly summarized as follows: Second-order beliefs about how own appearance and performance affect the subjective probability of being audited do change individuals' declaration decisions. In an environment in which the expected monetary payoff from cheating is positive (induced by a low penalty) the proportion of subjects who report honestly is higher if individuals know that their individual audit probability depends on the customs officer's decision. The share of subjects who honestly declare and pay duties increases roughly by 15-20 percentage points (compared to a purely random audit). In contrast, in an environment that induces individuals to report honestly (high penalty) there are no treatment effects. A large number of individuals seemingly believe that the customs officer would detect their dishonest declaration. In general, we can conclude that better compliance is induced if the audit depends on individuals' behavior when talking to the customs officer. These findings have policy relevance for the design of the set-up in which compliance decisions are made. A framework with personal contact and with discretion about whom to audit can improve compliance. This may hold not only for customs declarations but also for tax declarations and other compliance frameworks.

The problem of tax compliance more generally has generated enormous interest among economists.[2] One branch of this literature considers tax compliance as an incentive problem. Seminal papers in this context are Allingham and Sandmo (1972) and Yitzhaki (1974), Reinganum and Wilde (1985, 1986), and Chander and Wilde (1998). Apart from the purely monetary incentives, other aspects have been considered as possible determinants of compliance behavior. These include an intrinsic motivation (Frey 1997), an inclination for pro-social behavior (Frey and Torgler 2007), fairness considerations (Hartner et al. 2008), religiosity (Torgler 2006), and patriotism (Konrad and Qari 2011).[3] Feldman and Slemrod (2007) and especially Kleven et al. (2011), who analyze data from Danish tax authorities, find evidence that tax evasion is higher for self-reported income.[4] A considerable amount of experimental

[2] Much of this literature assumes away the complexities of tax declarations and reduces the problem to a compliance decision similar to the decision of travelers at the customs. Theory contributions focusing on customs compliance more explicitly are Thursby et al. (1991), and Yaniv (2010).

[3] Andreoni et al. (1998) and Slemrod (2007) provide in-depth surveys of this large literature.

[4] Alm et al. (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. Alm and co-authors base their argument on experimental evidence that more service for taxpayers from the tax administration leads to more honest behavior. 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

literature has analyzed the tax compliance decision in detail. Alm (2010, p. 654) reports that a considerable 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 related to the material-rewards-oriented decision models on tax compliance. He also surveys the experimental literature that considers social norms (e.g., groups' willingness to tolerate tax evasion, attitudes vis-à-vis the government, country-specific effects), and the issue of simplicity versus complexity. We consider the possible role of subjects' perceptions about whether they can influence the beliefs of others about their honesty in a compliance situation. We find that such second-order beliefs play a role, and this has implications for the optimal institutional design of compliance situations.[5]

The underlying theory to our experimental analysis about the power of imagination has also some links with the theory of beliefs about beliefs. For their compliance decision subjects need to form a belief about the likelihood of being audited. In two of our treatments the audit probabilities are known to be objective probabilities, and independent of individuals' behavior, even though they have face-to-face contact with the customs officer in one of the treatments. In the third treatment, in which they have face-to-face contact with a customs officer, this face-to-face contact has an influence on whether this subject is audited. Accordingly, their belief about whether they will be audited is a belief about the customs officer's belief, and about how they may be able to affect this officer's belief. We study these beliefs by studying the subjects' actions, namely their compliance decision.[6] Investigating the underlying reasons for changes in behavior is a matter of psychological game theory as established by Geanakoplos et al. (1989) for static models and Battigalli and Dufwenberg (2009) for dynamic models. From a behavioral perspective, in this context a more compliant behavior can be driven by guilt aversion (Charness and Dufwenberg 2006, among others), costs of lying (Vanberg 2008) or other traces of human psychology. In our framework we do not study these micro-motives. Instead we focus on the distinction between personal, face-to-face contact and a more anonymous declaration via the computer, and for the case with face-to-face contact, between situations in which subjects cannot affect their own probability of being audited, and situations in which they can affect their own audit probability by their performance, keeping the aggregate audit probability

behavior. Any dishonest behavior in our setting is an intended strategic behavior based on second-order beliefs.

[5] The process of belief formation in a signaling game need not stop short at second-order beliefs. However, Arad and Rubinstein (2011) show that subjects generally do not use more than three steps of reasoning. Hence, second-order beliefs are in fact of practical relevance for individual behavior.

[6] Note that we study a customs compliance situation in an environment that uses elements of a field experiment. In the course of the experiment, subjects have to walk to a neighboring room and talk to a customs officer. Since this 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 et al. (2004). See Manski and Neri (2011) and Costa-Gomes and Weizsäcker (2008) for examples and discussion of the more complex approach of eliciting beliefs via scoring rules.

constant.[7]

In Section 2 we explain the theoretical model and the experimental design. In Section 3 we derive testable hypotheses. These are given a more formal underpinning in the Appendix. Section 4 shows our findings about the power of imagination. Section 5 concludes.

2. The theoretical and experimental setup

We consider the following compliance decision. A traveling person i has either a high endowment (x_H) or a low endowment (x_L) of goods, with $x_L < x_H$, and the person knows what this own endowment $x_i \in \{x_L, x_H\}$ is. Customs knows that x_i is a random draw from the set $\{x_L, x_H\}$, and that with probability 0.8 the person i has the high value x_H, with probability 0.2 person i has the low value x_L, that is, customs knows the distribution from which x_i is drawn, but cannot observe i's endowment directly. At the customs i must declare 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 or not. We denote the two alternatives $a_i \in \{0, 1\}$. For $a_i = 0$ subject i can pass without audit, and for $a_i = 1$ the subject i receives an audit. The audit perfectly reveals the person's true endowment. The following payoffs for person i are (exogenously) attached to the different combinations of actions for the different endowments:

$$\begin{aligned}
\pi_i(x_L, l, 0) = \pi_i(x_L, l, 1) &= x_L \\
\pi_i(x_L, h, 0) = \pi_i(x_L, h, 1) &= x_L - T \\
\pi_i(x_H, l, 0) &= x_H \\
\pi_i(x_H, l, 1) &= x_H - T - \theta \\
\pi_i(x_H, h, 0) = \pi_i(x_H, h, 1) &= x_H - T
\end{aligned}$$

These payoffs conform with the intuitive outcomes: low-endowment persons pay no duties if they report truthfully, regardless of whether they receive an audit. If, for whatever reason a player reports a high endowment, the person has to pay a duty equal to T, also regardless whether an audit occurs or not. Persons with high-endowment who report truthfully have to pay a duty equal to T, regardless of whether they receive an audit. Persons with high-endowment who declare a low endowment receive different payoffs dependent on whether they receive an audit or not. If a person is not audited, no duty and no fine is to be paid. If a high-endowment person who reported a low endowment receives an audit, the person has to

[7] We derive our testable hypotheses on the basis of utility-maximizing behavior. As the role of second-order beliefs for compliance has not been analyzed yet, this paper will contribute to two out of Roth's three famous categories of experiments (Roth 1995). First, starting from our hypotheses we are "searching for facts" whether or not second-order beliefs induce strategic compliant behavior. And second, we are "whispering in the ears of princes" with the policy implications of our experimental results.

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 persons who maximize their payoff report truthfully. A risk-neutral high-endowment person i who maximizes own expected monetary payoff prefers to report truthfully if

$$x_H - T > p_i(x_H - T - \theta) + (1 - p_i)x_H,$$

where p_i is the probability that i attributes to being audited in case of declaring l.[8] A person who cares about monetary incentives only should therefore be indifferent between compliance and non-compliance if

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

Much of the further analysis is affected by how p_i is determined. Inspired by the insights from our survey among real customs officers, we consider three different treatments. Throughout all treatments, care is taken that players do not exchange views, and do not learn about other subjects' monetary payoffs in the end of the experiment when payments are made.

In the fully computerized baseline treatment (T1) the person i learns the value of own endowment while sitting in front of a computer. The person is then asked on the computer screen whether to declare high or low endowment: $y_i \in \{h, l\}$. More specifically, we ask whether i has to declare a high endowment. The person knows (as this is written down as part of the instructions) that the computer chooses $p_i \equiv 0.5$ if $y_i = l$. The subjects make these decisions in a laboratory room in which 20 subjects perform the same task. But as $p_i \equiv 0.5$ is given exogenously, each subject's task is formally independent of the tasks and choices of other subjects. In T1, for $p_i \equiv 0.5$, the indifference condition (1) reduces to $T = \theta$. The own-material-interest prediction is that a subject should choose $y_i = l$ if $T > \theta$ and $y_i = h$ if $T < \theta$.

In treatment T2, the subjects first learn their endowment while sitting in front of a computer in the same laboratory room as in T1. The 20 subjects waited until they were asked sequentially to walk into one out of two separate neighboring rooms. The sequencing of subjects was determined randomly.[9] In each of the two rooms a customs officer waited for subjects and saw the identification number of the subject entering on a computer screen.[10] The customs officer

[8] 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 does not have explanatory power in our data (see also below for details).

[9] By design the experiment precluded subjects from meeting each other during the compliance procedure or from inferring preceded or succeeded them in their room. All players 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. Jointly with the instructions participants had received a map showing the position of the two rooms in question, and signs directed them also to the respective rooms.

[10] To facilitate the comparison of the compliance behavior across treatments, the communication between the

first confirmed the entrant's identification number. Then, the subject had to report y_i, that is, whether he or she had something to declare. The officer entered the subject's report in the computer. The subject returned to the laboratory room. The first round was over after ten subjects had reported to one officer and the other ten subjects to the other officer. The officer in one room was female, the other officer was male, in each round. Also in T2 the person knows (as this is written as part of the instructions) that $p_i \equiv 0.5$, and independent of the person's (or other persons') general appearance or performance, i.e. the customs officer has no active decision role. This treatment takes into account that it may make a difference if persons have to report to a person rather than to a machine, as their subjective cost of misreporting need not be the same in both situations.[11] Therefore, we conducted the control treatment T2 in which compliance behavior different than in T1 can be traced back to a pure face-to-face effect.

A third treatment T3 is similar in structure to treatment T2. The only difference is how p_i is determined. In T3, subjects know that p_i is not exogenous. Instead, the customs officer has to assess all subjects and their declarations. The officer ranks these subjects according to his beliefs about whether they are cheaters.[12] This ranking enters into a procedure that determines the subjects who receive an audit, increasing the likelihood for an audit for subjects whom the officer considers more likely to be dishonest, and decreasing the likelihood for an audit for subjects whom the officer considers to be more likely to be honest. We construct a method by which these assessments enter into the audit decision, but by which, in the aggregate, precisely half of the subjects of type (x_H, l) receive an audit (see the Appendix for details). This procedure ensures that the aggregate audit probability is constant across treatments for all potential cheaters thereby removing a potential confounder of the treatment effects. These rules are common knowledge. Consider now a subject i; the subjective audit probability p_i is lower than 0.5 if subject i expects to be assessed as being more likely to be honest than the median among all other subjects who choose to underreport, and p_i 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 subjects who choose to underreport. Subjects must therefore form a belief about how their appearance and performance affects the beliefs of the customs officer. Subjects consider their audit probability to be lower than the audit rate of 0.5 that holds in the aggregate if they

customs officer and the subjects has to be standardized. More important, it is necessary to ensure anonymity for the subjects. For these reasons, customs officers were not recruited from the pool of student subjects. Instead, young employees and contract workers from the Max Planck Institute played the role of customs officers.

[11] 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.

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

believe that the officer believes them to be honest more than he believes other untruthfully reporting subjects to be honest. Their second-order beliefs may lead them to comply less truthfully in T3 than in T1 and T2. We will refer to this kind of subjects as confident or "strong" liars. In contrast, subjects who believe that the customs officer is more likely to rate them as a liar than other underreporting subjects, have a subjective audit probability p_i that is higher than 0.5. For these individuals their second-order beliefs can result in a more compliant behavior in T3 than in T1 or T2. Individuals of this type will be called inconfident, or "weak" liars.

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 in comparison to 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. In order to address this problem, we conducted all treatments for two different parameter settings. One of the settings was chosen to elicit and measure the difference in behavior of strong liars between T2 and T3. The second setting was chosen to elicit and measure the difference in behavior of weak liars between T2 and T3. We explain this further below in more detail.

We conducted the experiment in the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) in 2011. A number of twelve sessions took place. The subjects were students of diverse fields at the University of Munich.[13] In each session there were 20 subjects, totaling a number of 240 subjects who participated in the role as travelers. Each session started with a reading of the instructions, which were also distributed in written format. 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. In this way we ensured that the participants fully understood the nature of the experiment, especially the audit mechanism. Then the actual experiment started. Each individual participated only in 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 independent rounds. Independence was induced by a replacement of the person who served as the customs officer between the rounds. After the experiment, the subjects had to answer a questionnaire. We asked them for their gender, their field of study, etc. This quiz was followed by a standard risk elicitation game in the style of Holt and Laury (2002). Participants were asked to compare ten pairs of lotteries sequentially. This game also

[13]The participants were recruited using the software ORSEE (Greiner 2004).

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. In the end they received their earnings from the experiment plus the outcome of that lottery, plus a show-up fee of 4 euros.

Whereas customs officers, in line with reality, were paid flat, subjects participating in the compliance game generated their earnings as follows. At the beginning of each round of the experiment the subjects were sitting in front of their computer in the MELESSA experimental laboratory room. The value of the goods to be imported by them (their "endowment") was determined and displayed to them on their screens. The value was private information and it was either high (with $x_H = 1000$) or low (with $x_L = 400$). Then, each subject chose whether to report h(high) or l(low). If they reported $y_i = h$, they had to pay customs duties of $T = 200$. If a subject with a high value reported l and was audited, the subject had to pay customs duties of $T = 200$ plus a fine θ. For θ we used two different parameter settings. In half of all sessions the fine was equal to $\theta = 100$, in the other half of the sessions the fine was equal to $\theta = 300$. The audits were carried out by the computer and if an audit took place, the true value of x_i was found. Subjects learned whether they received an audit at the end of each round, and the resulting monetary payoff was shown to them on the computer screen. The participants' earnings from the experiment consisted of the outcome of one specific round that was randomly drawn by the computer. The currency in the laboratory was named talers.[14]

3. Hypotheses

The experiment addresses the role of individuals' subjective beliefs about their own ability to influence customs officers. Such beliefs can be formed in a compliance situation with direct face-to-face communication with customs officers who have an influence on who receives an audit. In a fully anonymous, computerized compliance framework with exogenous audit probability such beliefs cannot play a role. We test whether there are subjects who think that they are "weak liars" who would not declare truthfully in T1 or T2, but prefer to declare truthfully in an environment such as T3, and whether there are subjects who think that they are "strong liars", who would declare truthfully in T1 or T2, but prefer to declare untruthfully in an environment such as T3. 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.

[14] In the treatment T1, 1000 talers were converted into 10 euros. In T2 and T3, 1000 talers were converted into 15 euros. 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 the treatments T2 and T3 lasted longer.

Consider the indifference condition (1) for θ for players who are motivated by their monetary payoffs. This condition yields the prediction that all players with $x_i = 1000$ declare honestly for $\theta = 300$, whereas all players with $x_i = 1000$ declare $y_i = l$ for $\theta = 100$, for both treatments T1 and T2. The difference in punishment size should have qualitatively a similar effect in T3. In the experiment, we expect that other idiosyncratic factors may also play a role in an individual's compliance decision. We do therefore not expect these predictions to materialize sharply. However, we would expect that, among the subjects with $x_i = 1000$, the share of subjects who declare honestly is higher for $\theta = 300$ than for $\theta = 100$.[15] These high-fine and low-fine treatments constitute the benchmark case that is to be compared with a treatment T3 with high fines and with low fines.

Before turning to the comparison of the baseline treatment with T3, we test for the role of face-to-face contact of the compliance decision. For players who are motivated by the monetary payoffs, the declaration mode (face-to-face or automated) should not matter, as long as the audit probability is unaffected. We formulate this as an auxiliary hypothesis: *There is no treatment effect between T1 and T2 for subjects with* $x_i = 1000$. This hypothesis is derived more formally in the appendix as Proposition 1. It is auxiliary to what we do, as it paves the ground for our main research question.[16]

We formally develop our central hypotheses for a comparison between T1 and T3 in the Appendix. The hypotheses follow directly from the characterization of the Bayesian Nash equilibrium in Proposition 2 in this appendix. The formal framework assumes that individuals differ in exogenous appearance characteristics. We aggregate these characteristics in what we label a subject's *look of being honest.* We do not consider what this look means precisely, and whether this is an objective feature that somehow forces the subject to be more honest in equilibrium than others, or whether this quality exists only in the imagination of the subjects. Suppose that subjects cannot alter their own look, and that this look is the quasi automated

[15] 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, as has been shown in a number of other experiments.

[16] Nevertheless, it was unclear ex-ante whether this hypothesis holds. 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 subjects with $x_i = 1000$ who declare honestly is higher in T2. Several reasons for such an alternative outcome could be considered. For instance, the possible cost of lying may be higher if lying occurs in a situation with face-to-face contact. Also, subjects may be too much used to the idea that the information generated by face-to-face contact is used for determining the individual audit probability. They may be unable to abstract completely from this effect in their decision-making even in a situation in which, objectively speaking, face-to-face contact has no effect on their probability of being audited. If the compliance behavior changed substantially between T1 and T2, this would suggest that a change in the compliance behavior from T1 and T3 is driven by a mixture of a mere face-to-face effect. If that is the case, the pure 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 second-order beliefs.

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 players who actually underreport. By construction, the customs officer does not consistently solve for an equilibrium, but simply sorts subjects according to their looks and half of the subjects in this subset are automatically subject to an audit. This half is, by construction, the less-honest-looking half in the eyes of the customs officer.[17]

For the numerical case with high fines ($x_H = 1000\,; T = 200; \theta = 300$), the critical level of p_i for which i is indifferent about whether to report truthfully is $p_i = 2/5$ by condition (1). Accordingly, the equilibrium prediction for material-payoff-motivated subjects is that some (particularly honest-looking) subjects will underreport in the treatment T3 with high fines, whereas, for the same high fines, the prediction for T1 and T2 was that no subject should underreport for this parameter range. This yields our first main hypothesis.

Hypothesis A: *For $\theta = 300$, the share of players $x_i = x_H$ 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 for which the subject is indifferent about reporting truthfully or not is $p_i = 2/3$ by condition (1). Accordingly, the equilibrium prediction for material-payoff-motivated subjects is that some (particularly dishonest-looking) subjects will report truthfully in the 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 fines. This yields our second main hypothesis.

Hypothesis B: *For $\theta = 100$, the share of players with $x_i = x_H$ who declare honestly in T3 is larger than in T1 or T2.*

Hypotheses A and B establish our main testable hypotheses about possible compliance-*decreasing* effects and compliance-*increasing* effects of second-order beliefs in the two penalty regimes, respectively. We now turn to the data and results.

4. Results

In this section we analyze subjects' aggregate behavior across treatments using different econometric models. We first describe briefly the characteristics of our sample and the associated

[17] We instructed the customs officers to "assess all subjects" and to "grade the honesty of their declarations" at the end of the round. Since each officer met each traveler just once, those two assessments are likely to coincide in practice.

empirical strategy. We then estimate treatment effects according to the between-subjects design outlined in the previous section.

4.1. Sample Characteristics and Empirical Strategy

As described in section 2, each participant in T1 played ten rounds, while the number of rounds equals four in T2 and T3. Each treatment is carried out twice for both setups with respect to the penalty. In the first setup we have $T > \theta$ such that dishonest behavior is induced, while the second setup ($T < \theta$) induces incentives to honestly report a high endowment. 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 corresponding 640 observations from T2 and T3. Note that subjects were assigned their endowment in each round randomly with replacement. Hence, the number of low- and high-endowment observations respectively is not fixed ex-ante.[18]

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}	
		400	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 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 in the following only high endowment observations (where $x_{it} = 1000$).

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

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

regression equation reads

$$y_{it} = \mathbf{\Xi'_{it}}\boldsymbol{\beta} + u_i + \epsilon_{it} \tag{2}$$

where the binary variable y_{it} is equal to one if subject i who has 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. We employ linear and logistic mixed effects models (multi-level models) to fit equation (2).

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. Second, the number of high-endowment observations differs across subjects. As the probability for 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 a few subjects are observed once or twice. We therefore use the well-established parametric approach to handle such sampling conditions and employ linear and nonlinear (logistic) mixed effects models (e.g. Agresti 2003, Wooldridge 2006, Cameron and Miller 2009). The term "mixed effects model" refers to the fact that both fixed and random effects are estimated.[19]

We mainly present the results from a linear mixed model that essentially ignores that the explained variable is binary. However, a linear model is a useful starting point as the estimates are easy to interpret and are often in line with the results from probit and logit models. The robustness section presents the results from a corresponding logistic model (which takes into account that the explained variable is binary) that closely resemble the results from the linear model.

4.2. Treatment Effects

Table 2 and Figure 1 present the aggregate compliance rate by Treatment and penalty. Starting

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

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

[19] 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).

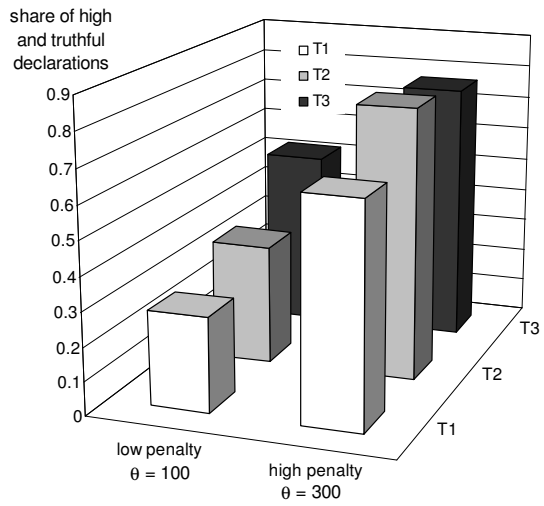

Figure 1: Reporting by Treatment and penalty

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 35% in T2. By contrast, the fraction of honest reports is close to 52% in T3. This suggests that the influence of mere face-to-face contact is much smaller than the effect of the endogenous subjective 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 a compliance increasing effect driven by subjects' imagination with respect to their audit probability. Turning to the high-penalty setup ($T < \theta$), no clear pattern emerges. Compared to the compliance rate in T1 (65%), 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 provides support for Hypothesis B while there is no support for Hypothesis A.

We now move to the econometric evidence and start with a linear probability model. To this end, we fit equation (2) by maximum likelihood assuming that both u_i and ϵ_{it} follow a normal distribution.[20] 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 estimates presented in the two tables confirm the descriptive evidence. In the low-penalty case (Table 3), the fraction of truthful reports in T1 equals 28%. Compared to

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

Table 3: Linear Unobserved Effects Model for $\theta = 100$ (low penalty)

	(1)		(2)	
(Intercept)	0.28	(0.06)	0.38	(0.04)
Treatment2	0.08	(0.08)		
Treatment3	0.23	(0.08)		
Log likelihood	-289.90		-293.69	
AIC	589.81		593.38	
Likelihood Ratio	7.569 (p=0.0223)			
Observations: 575, groups: 120				

Standard errors in parentheses.

The table presents the results from a linear probability model where the binary variable y_{it} (report) 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.

T1, the compliance rate in the intermediate treatment T2 is eight percentage points higher. However, the associated standard error of 0.08 reveals that this difference is not precisely estimated. The compliance rate in T3 is 23 percentage points higher compared to T1, yielding a total compliance 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 tax compliance in T3 is driven by individuals' imagination with respect to their own audit probability. 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 4: Linear Unobserved Effects Model for $\theta = 300$ (high penalty)

	(1)		(2)	
(Intercept)	0.65	(0.05)	0.73	(0.03)
Treatment2	0.14	(0.08)		
Treatment3	0.12	(0.08)		
Log likelihood	-259.20		-261.09	
AIC	528.40		528.18	
Likelihood Ratio	3.7797 (p=0.1511)			
Observations: 580, groups: 120				

Standard errors in parentheses.

The table presents the results from a linear probability model where the binary variable y_{it} (report) 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 around 12-14 percentage points higher in T2 and T3.

We can assess the 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. In the high penalty setup, the Akaike information criterion (AIC) of the model including treatment effects is equal to 528.40 (see Table 4, column (1)) and exceeds the corresponding value of the more parsimonious model. Hence, the more parsimonious model is preferred. By contrast, in the low-penalty setup the model including treatment dummies (AIC 589.81, see Table 3, column (1)) is preferred over the more parsimonious model (AIC 593.38, column (2)). A comparison of the associated likelihood ratios supports the same conclusions. Summarizing, the preferred model in the high-penalty case is a model without treatment dummies and an average compliance rate of approximately 73% (Table 4, column (2)). Hence, the estimation results for the high-penalty case do not support hypothesis A.

4.3. Robustness Checks

4.3.1. Overview

This section inquires the robustness of the estimated treatment effects in a number of ways. First, we check whether the use of a linear probability model has produced biased coefficients. We use the common approach for binary response variables and consider logistic models. Note that there are two main approaches for estimating such models. The first approach is a generalized linear mixed model assuming that the subject-specific error term follows a normal distribution. This model produces coefficients that are suitable for predicting the probability of the binary compliance variable conditional on the random intercept u_i. For this reason, it is sometimes referred to as the "subject-specific-model", especially in the biostatistics literature (see, for example, Agresti 2003). In the economics literature, this model is usually referred to as "random effects logit model" (e.g. Wooldridge 2010, Chapter 15). Unconditional marginal effects ("population-averaged effects") can be obtained by integrating over the distribution of the estimated subject-specific random effect.

The second approach treats the unobserved heterogeneity as nuisance parameters and models directly the marginal mean (population mean) of the response variable. In our case, both approaches produce the same evidence with respect to the marginal (population-averaged) effects. The marginal effects are the quantities of interest, as our experiment is designed to exploit between-subject variation rather than within-subject variation. We therefore present

only the results from the marginal model.[21]

4.3.2. Baseline Results

The first column of Table 5 and Table 6 summarizes the baseline results, where the probability

Table 5: Marginal Logistic Unobserved Effects Models (low penalty)

	(1)	(2)	(3)	(4)	(5)
(Intercept)	-0.943	-0.883	-0.545	-1.123	-1.301
	(0.124)	(0.152)	(0.253)	(0.200)	(0.319)
Treatment2	0.345	0.325	0.114	0.342	0.360
	(0.225)	(0.226)	(0.255)	(0.227)	(0.229)
Treatment3	1.035	1.034	0.825	1.031	1.039
	(0.215)	(0.215)	(0.243)	(0.216)	(0.217)
RiskMeasure		0.0407	0.0390	0.0430	0.0390
		(0.0476)	(0.0477)	(0.0474)	(0.0478)
Second Half		-0.115		-0.121	-0.121
		(0.179)		(0.179)	(0.179)
round			-0.0734		
			(0.0419)		
Female				0.364	0.374
				(0.191)	(0.192)
Age					0.0223
					(0.0311)
Observations: 575, groups: 120					

Standard errors in parentheses.

The table presents the results from a marginal logistic unobserved effects model where the binary variable y_{it} (report) is predicted by a set of treatment dummies and control variables (fixed effects). The subject-specific error terms are treated as nuisance. The omitted reference category is Treatment 1. The model is fitted using generalized estimating equations (GEE).

of reporting truthfully a high endowment is predicted by a subject-specific error term and a set of treatment dummies. The estimated coefficients are comparable to the linear probability model from the previous section. For example, the predicted compliance rate in Treatment 1 for the low-penalty case is given by $\Lambda(-0.943) = 0.28$ where $\Lambda(\cdot)$ denotes the inverse logit function. The respective calculations for the other treatments show that the estimated treatment effects for T2 and T3 are 0.07 and 0.23 respectively. Hence, the estimates are very close to the results from the linear model. The same holds for the high-penalty case (Table 6, column 1).

4.3.3. Variation over time and risk attitudes

Figures 2 and 3 depict the compliance rate in each round averaging over subjects for the low-

[21] The marginal model is fitted using generalized estimation equations (GEE).

Table 6: Marginal Logistic Unobserved Effects Models (high penalty)

	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.637	0.807	1.091	0.804	0.199
	(0.118)	(0.156)	(0.258)	(0.200)	(0.302)
Treatment2	0.768	0.577	0.362	0.577	0.625
	(0.249)	(0.256)	(0.284)	(0.257)	(0.259)
Treatment3	0.567	0.564	0.354	0.565	0.627
	(0.239)	(0.243)	(0.273)	(0.244)	(0.248)
RiskMeasure		0.238	0.240	0.238	0.249
		(0.0527)	(0.0529)	(0.0532)	(0.0537)
Second Half		-0.204		-0.204	-0.192
		(0.190)		(0.190)	(0.191)
round			-0.0701		
			(0.0405)		
Female				0.00364	0.00980
				(0.196)	(0.197)
Age					0.0716
					(0.0273)
Observations: 580, groups: 120					

Standard errors in parentheses.

The table presents the results from a marginal logistic unobserved effects model where the binary variable y_{it} (report) is predicted by a set of treatment dummies and control variables (fixed effects). The subject-specific error terms are treated as nuisance. The omitted reference category is Treatment 1. The model is fitted using generalized estimating equations (GEE).

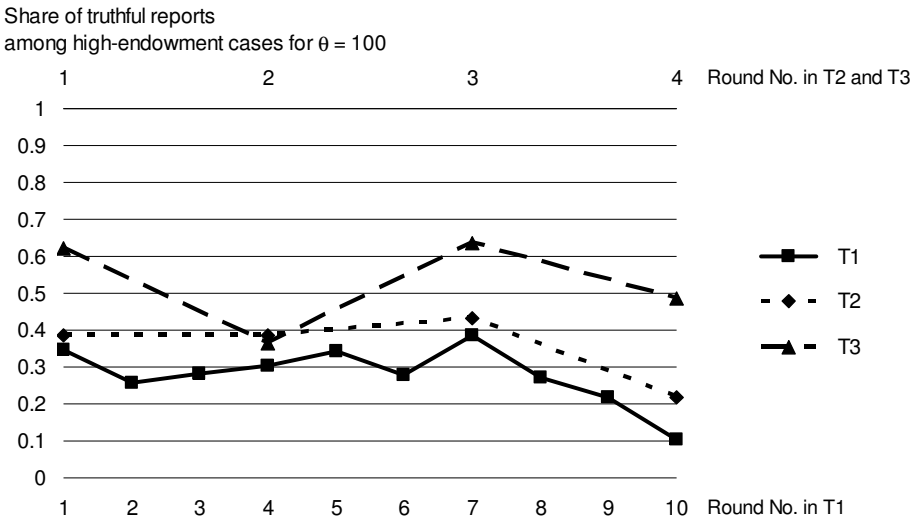

Figure 2: Reporting behavior among high-endowment cases for Treatments T1 (black), T2 (dotted) and T3 (dashed) for $\theta = 100$. Note that the fully computerized treatment T1 had 10 rounds whereas the treatments T2 and T3 with personal interviews had only four rounds, but each of the rounds in T2 and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

Figure 2: Reporting by Treatment and rounds (low penalty)

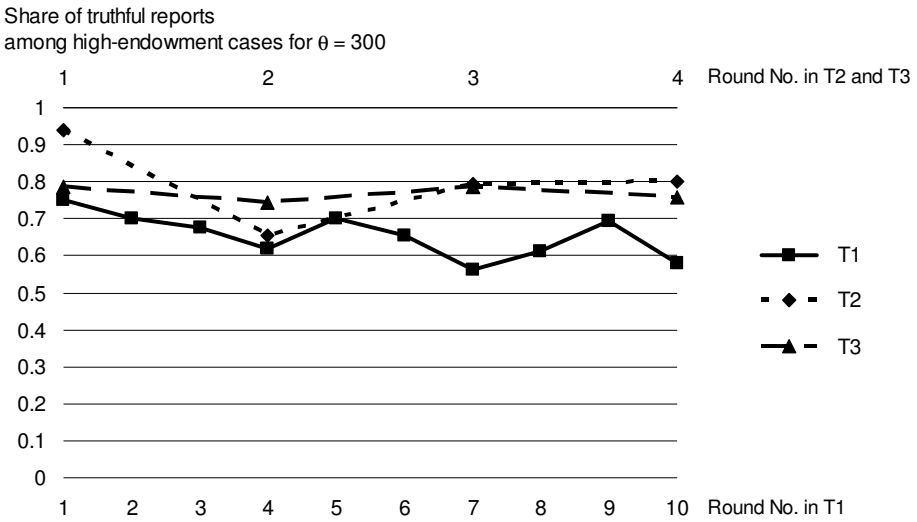

Figure 3: Reporting behavior among high-endowment cases for Treatments T1 (black), T2 (dotted) and T3 (dashed) for θ = 300. Note that the fully computerized treatment T1 had 10 rounds whereas the treatments T2 and T3 with personal interviews had only four rounds, but each of the rounds in T2 and T3 took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

Figure 3: Reporting by Treatment and rounds (high penalty)

and high-penalty setup respectively. Especially for the latter case, there is some evidence that the compliance rate decreases in later periods. To facilitate a comparison across the treatments, column (2) of Tables 5 and 6 respectively enters a dummy variable indicating the second half of the game into the regression. We also introduce the individual measure of risk aversion into the model.

The estimated treatment effects in the low-penalty setup (Table 5) are robust 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 6) obtains a value of 0.238. As expected from the figures, the estimated time coefficient is larger compared to the low-penalty case. However, the associated standard error is rather large.

Column (3) of the two tables considers a linear time trend instead of the dummy variable. Note that the coefficient in this specification 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.

4.3.4. Extended set of control variables

Column (4) and (5) of tables 5 and 6 introduce a gender dummy and an age effect, respectively, as additional control variables. The gender coefficient in the low-penalty setup (Table 5) obtains values around 0.364. This is a sizable effect comparable to an increase in the compliance rate of seven percentage points. However, the standard error of 0.192 indicates that the precision of this estimate is rather poor. The age coefficient (column 5) is also noisy and small. In the high-penalty setup (Table 6) the gender and the age coefficients are small and imprecise. More important, the results obtained so far are robust in both penalty setups.

4.3.5. Using T2 as the reference treatment

Note that all 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 5 and 6 respectively, there is no direct estimate for the difference between T3 and T2. We therefore run a marginal logit model where the linear predictor is transformed to reflect the difference between adjacent treatments.

Table 7 summarizes the results. The first column shows the coefficients for $\theta = 100$. The

Table 7: Marginal Logistic Unobserved Effects Model – difference contrasts

	(1) $\theta = 100$	(2) $\theta = 300$
(Intercept)	-0.483	1.082
	(0.0952)	(0.108)
T2-T1	0.345	0.768
	(0.225)	(0.249)
T3-T2	0.690	-0.201
	(0.257)	(0.302)
N	575	580

Standard errors in parentheses.

The table presents the results from a marginal logistic unobserved effects model where the binary variable y_{it} (report) is predicted by a set of treatment dummies and control variables (fixed effects). The subject-specific error terms are treated as nuisance. The coefficients show the difference between adjacent treatments. The model is fitted using generalized estimating equations (GEE).

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 (eight 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 shows a value of roughly 2.68. Column (2) presents the coefficients for $\theta = 300$. In line with previous results, the estimates

suggest that the aggregate compliance rates in T2 and T3 are similar and higher compared to T1.

4.3.6. Robustness Checks: Summary

Summarizing, all models corroborate the evidence from the linear model: 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 an compliance-increasing effect, driven by subjects' imagination with respect to their own audit probability. For $\theta = 300$ the treatment dummies do not explain much variation. Recall that all results are obtained from unobserved effects models to control for heterogeneity of subjects.

5. Conclusion

We analyzed the role of imagination for subjects in a customs compliance framework. Our experimental results reveal a major asymmetry: A considerable number of subjects behave as if they consider their deceptive ability to be very low, but there is no evidence of subjects who behave as if they consider their deceptive ability to be very high.

In a first set of experiments the subjects decide in compliance frameworks in which they have monetary incentives to underreport, due to low fines. In one treatment customs officers make assessments on the basis of personal communication with face-to-face contact. In this treatment their appearance or performance influences which subjects receive an audit. Aggregate compliance behavior is substantially higher (15-20 percentage points) in this treatment compared to a treatment with a strictly random audit. Using a further control treatment we can also distinguish between the role of face-to-face contact and the role of customs officers' assessment and the formation of subjective beliefs. We find that higher compliance is driven by subjective beliefs rather than by face-to-face contact.

In a second set of treatments we provide subjects with monetary incentives to report truthfully, due to high fines. In such an environment only those subjects would have incentives to underreport who believe that they can successfully fool the customs officer. We do not find evidence for such behavior.

In interviews with customs officers from three international German airports, about 50% indicated that they consider face-to-face interviews as the most effective strategy to increase customs compliance. Our experimental evidence is in line with this perception.

Our findings about the power of imagination have policy relevance not only for the institutional framework of customs declarations. The results are also suggestive for 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.

A. Appendix

In this appendix we consider the compliance decision of players with exogenous and with subjective audit probabilities as a Bayesian game between the subjects with high income, with other subjects with low income also being present. The framework is as follows. There is a set I of n players i with a high endowment x_H. Also, there are m players who have a low endowment x_L. All players simultaneously make a reporting decision. The report is $y_i \in \{l, h\}$. Players who declare low endowment may be subject to an audit. For the sake of simplicity, we assume that the audit itself is not costly for the player who is subject to an audit. The audit reveals the true endowment of the player with certainty.

Players with low endowment have a strictly dominant strategy: to report low income. Therefore we truncate their problem here, assuming that they follow this strategy and always report truthfully. A player i with high endowment either declares high income ($y_i = h$) and pays a tax $T(= 200)$ and receives a monetary income of $x_H - T$ or i declares low income ($y_i = l$). If a player with high endowment is audited, the underreporting is detected and the player has to pay the tax T and a fine θ. If the player is not audited, the player enjoys the full income x_H.

The treatments we considered in the main part of the paper had different mechanisms determining whether a player who declares $y_i = l$ receives an audit. In T1 and in the control treatment T2, there was an exogenously given probability for being audited. This probability was denoted p_i. With the complementary probability $(1 - p_i)$, no audit occurs. Accordingly, the expected monetary payoff for player i in T1 and T2 is $p_i(x_H - T - \theta) + (1 - p_i)x_H$. For this exogenous audit probability the decisions of all subjects are fully independent. Subjects with high endowment choose to report truthfully if $x_H - T \geq p(x_H - T - \theta) + (1 - p)x_H - \Delta_i$, or, for $p_i \equiv p = 1/2$, they choose to report truthfully if

$$pT + p\theta > T. \tag{3}$$

In the treatments T1 and T2, this condition was fulfilled for the high-fine case with $\theta = 300$ and the reverse condition was fulfilled for the low-fine case with $\theta = 100$. Even though this is an almost trivial decision problem with a trivial prediction, it constitutes the status of reference for the game underlying treatment T3, and this is why we state this as a proposition.

Proposition 1: *In T1 and T2, if players are motivated by monetary incentives only, then, in the equilibrium, all players with high income report truthfully if* $\theta = 300$ *and underreport if* $\theta = 100$.

Proposition 1 leads to our auxiliary hypothesis in the main part of the paper. Note that we do not expect all players to behave strictly in accordance with this prediction. There are potentially many behavioral reasons why other considerations sometimes dominate the monetary incentives that drive the behavior in Proposition 1. In particular, some subjects may be more honest than is predicted by Proposition 1, for instance, because they feel a mental cost of lying (Lundquist et al. 2009), or they may feel good by paying taxes (Harbaugh et al. 2007). Some subjects may also underreport even if the monetary incentives suggest that they report truthfully, for instance, because they may like to gamble, or because they may enjoy lying. However, if these other attitudes do not interact systematically with the monetary incentives and if the subjects are all random draws from the same population with these underlying characteristics, we should observe that more subjects report truthfully if $\theta = 300$ than if $\theta = 100$.

We now turn to the more complex framework underlying T3. We first give a detailed account of this treatment. Recall that there are n subjects with high endowment and m subjects with low endowment. The latter all truthfully report low endowment. The former make a choice. Accordingly, the reporting decisions lead to a set of players characterized by their endowments and their reports. This set is $\{(x_H; y_1), ..., (x_H; y_n), (x_L; l), ..., (x_L, l)\}$. The customs officer has face-to-face communication with the subjects from this set and observes their reports. A subset of this set is the set of players who report low income. This subset includes all m subjects with low endowments, and a subset of the n subjects with high endowments. The customs officer solves the task of ranking the subjects. Let $(r_1, ..., r_{n+m})$ be the ranking emerging from this, with r_1 being the subject to whom the customs officer attributes the highest likelihood of being dishonest, and r_{n+m} the subject which looks most honest among the individuals who reported $y_i = l$. At this point the computer takes over. It uses the ranking to determine which subject from this set receives an audit. For this purpose, the computer "cleans" this ranking from all subjects who, in fact, have low endowments, by dropping these entries from the list, but preserving the ranking among the subjects who remain on this list. This leads to a reduced list $(\hat{r}_1, ..., \hat{r}_{k_n})$ which consists of k_n entries. This reduced list ranks all underreporting subjects with high endowment. The highest entry $\hat{r}_1$ in this list is the subject $(x_H; y_i = l)$ which in the eyes of the customs officer looks most suspicious within this subset. Having computed this list, the subjects on the upper half of this list receive an audit. If k_n is an even number, then the subjects $\hat{r}_1, ..., \hat{r}_{k_n/2}$ on this list are audited, their underreporting is detected, and they pay a fine equal to θ. If k_n is an odd number, then the $((k_n-1)/2)$ subjects $\hat{r}_1, ..., \hat{r}_{(k_n-1)/2}$ are audited with probability 1, and the subject $\hat{r}_{(k_n+1)/2}$ receives an audit with an exogenous probability of 1/2. Hence, this procedure makes sure that half of all subjects who are underreporting receive an audit, are detected and fined. Moreover, this procedure makes sure that the assessment of the customs officer sorts all individuals such that the subset of underreporting subjects who

receive an audit consists of those subjects whom the customs officer ranks as being more likely to be underreporting.

For a formal analysis we need to describe what drives players' beliefs about their deceptive abilities. We assume that the deceptive ability of individuals is an objective characteristic of subjects, and subjects know their own characteristic, and they know the general distribution of this characteristic across the population. Let $\lambda_i \in [0,1]$ measure the player's objective ability as a liar. That is, a player with $\lambda_i = 0$ has the lowest possible ability, and a player with $\lambda_i = 1$ has the highest possible ability. For each player i, let λ_i be drawn from the same distribution with cumulative distribution function $F(\lambda)$, assuming that $F(\cdot)$ is continuously differentiable in the interior of $(0,1)$, has full support and has no mass points. Let the customs officer also observe the individual value of λ_i for each player i who reports low income, and let the officer rank the players in an increasing order of their lambdas.[22] We can then state the following property:

Proposition 2: *Let there be more than one player who has low endowment. For a given θ, there exists a critical λ^* such that all players i with $x_i = x_H$ with $\lambda_i \leq \lambda^*$ report truthfully and all players i with $x_i = x_H$ with $\lambda_i > \lambda^*$ underreport.*

For a proof, consider player i who thinks that all other players follow this equilibrium strategy and confirm that the behavior that is described in Proposition 2 is a best response. Given the beliefs about other players' behavior as a function of their λ_j-values and knowing the distribution from which theses values are drawn, and given player i's own λ_i, the player can assess the probability of being audited if the player chooses to underreport. Recall that m players j who do have low endowment always report $y_j = l$, irrespective of their λ_j. The probability that s of these players have a λ_j that is lower than λ^* is equal to

$$\begin{pmatrix} m \\ s \end{pmatrix} F(\lambda^*)^s (1 - F(\lambda^*))^{m-s}. \tag{4}$$

Accordingly, the probability that at least k_m of these players have $\lambda_j \leq \lambda^*$ is equal to

$$\sum_{s=k_m}^{s=m} \begin{pmatrix} m \\ s \end{pmatrix} F(\lambda^*)^s (1 - F(\lambda^*))^{m-s}. \tag{5}$$

Among the $n-1$ players with high endowment other than player i, the probability that s of them underreport is equal to the probability that s of them have a $\lambda_j \geq \lambda^*$. This probability

[22] Note that, by design, the customs officer is not a fully rational player here, but simply ranks subjects according to their observed "honest look". The officer does not solve for how different λ feeds back into actual decision-making in a Bayesian Nash equilibrium with the tax officer as a player.

is equal to

$$\left(\begin{array}{c} n-1 \\ s \end{array} \right) F(\lambda^*)^{n-1-s}(1-F(\lambda^*))^s. \tag{6}$$

Accordingly, the probability that at most k_n other players have $\lambda_j \geq \lambda^*$ is equal to

$$\sum_{s=0}^{s=k_n} \left(\begin{array}{c} n-1 \\ s \end{array} \right) F(\lambda^*)^{n-1-s}(1-F(\lambda^*))^s. \tag{7}$$

Consider now a player whose $\lambda_i = \lambda^*$. If he underreports, he is not audited if $k_m > \frac{m+k_n+1}{2}$ and is audited with probability 1 if $k_m < \frac{m+k_n+1}{2}$. Note that for $\lambda^* = 0$ the probability that i with $\lambda_i = \lambda^*$ is audited is 1: by the assumptions about F, player i ranks at the bottom of the ranking of the customs officer with probability 1, and the bottom half of this list is audited. For $\lambda^* = 1$ the probability that i with $\lambda_i = \lambda^*$ is audited is zero for $m > 1$ and $n > 1$. To see this note that, for $\lambda^* = 1$, all m players who have low endowment report $y_j = l$; i.e., $k_m(\lambda^* = 1) = m$ with probability 1. Moreover, $k_n(\lambda^* = 1)$ is zero with probability 1, as i is the only player with such a high λ with probability 1. Accordingly, a player i with $\lambda_i = 1$ will end up at the top of the ranking. Consider now the effect of an increase in λ^* starting at some positive value λ^*. This increase continuously raises the probability that $k_m \geq s$ for any given $s = 1, 2, 3, ..., m$ in a monotonic and continuous manner. For given k_n it increases the probability that the condition $k_m > \frac{m+k_n+1}{2}$ is fulfilled. This makes it more likely that a player with a given $\lambda_i = \lambda^*$ who reports low income is not audited, for any given number of k_n. Further, an increase in λ^* reduces the number of underreporting players for any given draw $(\lambda_1, ..., \lambda_n)$ of abilities for the set of $n-1$ high-endowment players other than player i. Hence, for any given k_m, an increase in λ^* makes it more likely that the condition $k_m > \frac{m+k_n+1}{2}$ holds.

We now denote $p(\lambda_i; \lambda^*)$ the probability that a player with λ_i who underreports is audited if λ^* is the threshold as defined in Proposition 2. This probability is a continuous and (weakly) monotonically decreasing function in both its arguments, with $p(0,0) = 1$ and $p(1,1) = 0$. Accordingly, there is a λ^* and an induced $p^* = p(\lambda^*, \lambda^*)$ such that the condition

$$x_H - T = p^*(x_H - T - \theta) + (1-p^*)x_H \tag{8}$$

is fulfilled for one value of λ^*. For this threshold λ^*, high-endowment players with $\lambda_i = \lambda^*$ are just indifferent whether to report truthfully or to underreport. In turn, high-endowment players with $\lambda_i > \lambda^*$ prefer to underreport, and high-endowment players with $\lambda_i < \lambda^*$ prefer to overreport.

Our main hypotheses A and B follow now as a corollary of Proposition 2. For $\theta = 100$, all players with a high endowment who make a decision on the basis of their monetary incentives

choose to underreport in T1 and T2. For T3, the share of players who choose to underreport is smaller than 100 percent. Similarly, for $\theta = 300$, all players with a high endowment who make a decision on the basis of their monetary incentives choose to report truthfully in T1 and T2. For T3, there is a range of λ-types who report truthfully and a range of λ-types who underreport. Accordingly, in expectation, less than 100 percent of all players with high endowment choose to report truthfully.

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

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

Arad, A. and Rubinstein, A. (2011). The 11-20 money request game: A level-k reasoning study, mimeo.

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

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

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.

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.

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.

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.

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). Psychological games and sequential rationality, *Games and Economic Behavior* 1(1): 60–79.

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

Harbaugh, W., Mayr, U. and Burghart, D. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations, *Science* 316(5831): 1622–1625.

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

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. and Qari, S. (2011). The last refuge of a scoundrel? Patriotism and tax compliance, *Economica* (forthcoming).

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.

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, Economics Working Papers E00-279, University of California at Berkeley.

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.

Roth, A. (1995). Introduction to experimental economics, Vol. 1, Princeton, NJ: Princeton University Press, pp. 3–109.

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

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.

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

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, Vol. 1 of *MIT Press Books*, The MIT Press.

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.