Journal of Economic Behavior & Organization 124 (2016) 102–114

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

Do students behave like real taxpayers in the lab? Evidence from a real effort tax compliance experiment[☆]

C.Y. Lawrence Choo[a], Miguel A. Fonseca[b,*], Gareth D. Myles[c]

[a] University of Erlangen-Nuremberg, Germany
[b] University of Exeter, United Kingdom
[c] University of Exeter and Institute of Fiscal Studies, United Kingdom

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form
21 September 2015
Accepted 21 September 2015
Available online 3 October 2015

JEL classification:
C91
H26

Keywords:
Tax compliance
Real effort
Field experiment

ABSTRACT

We report on data from a real-effort tax compliance experiment using three subject pools: students, who do not pay income tax; company employees, whose income is reported by their employer; and self-employed taxpayers, who are responsible for filing and payment. While compliance behavior is unaffected by changes in the level of, or information about the audit probability, higher fines increase compliance. We find subject pool differences: self-assessed taxpayers are the most compliant, while students are the least compliant. Through a simple framing manipulation, we show that such differences are driven by norms of compliance from outside the lab.

> In this world nothing can be said to be certain, except death and taxes.
>
> (Benjamin Franklin, 1817)

1. Introduction

Tax is the primary tool used by governments to finance public administration and public services. However, due to the high costs of monitoring compliance, tax evasion is as old a concept as tax itself. Tax evasion remains an economically important problem in modern economies: the tax gap, which is the non-received tax revenue in a fiscal year, is estimated to be $450 billion in the United States in 2006 (IRS, 2012) and £35 billion in the United Kingdom in 2012 (HMRC, 2013).

The economic analysis of the tax compliance decision began with Allingham and Sandmo (1972) and Yitzhaki (1974). In this class of models, the taxpayer chooses the level of evasion which maximizes her expected utility, and risk arises from the possibility that a random audit may be conducted by the tax authority. The Allingham-Sandmo-Yitzhaki model predicts

☆ We are very grateful to Tim Miller for his outstanding help in programming the software and running the sessions. We are also grateful to Saros Research and ICM for their help in recruiting the non-student subjects. We also thank Kim Bloomquist, Tim Lohse, participants at the 2013 PET conference and the 2014 CEBID conference, as well as two referees and the editors for helpful comments and discussions. All remaining errors are ours alone. Financial support from HMRC and the ESRC/HMRC/HMT Tax Administration Research Centre (grant no. ES/K005944/1) is also gratefully acknowledged. The views in this report are the authors' own and do not necessarily reflect those of HMRC.

* Corresponding author. Tel.: +44 1392722584.
E-mail addresses: lawrence.cy.choo@fau.de (C.Y.L. Choo), m.a.fonseca@exeter.ac.uk (M.A. Fonseca), g.d.myles@exeter.ac.uk (G.D. Myles).

http://dx.doi.org/10.1016/j.jebo.2015.09.015

that tax evasion will fall when either the penalty rate or the probability of being caught evading increase. However, when confronted with values of the audit probability and the penalty rates close to those observed in practice, the model predicts that all taxpayers should evade. This is contradicted by evidence of generally high levels of compliance in most western economies: despite the large size of the estimated tax gap in the US, it only amounts to about 17% of total tax liabilities.[1]

The discrepancy between the predictions of the model and the data led some to argue that high levels of compliance are due to psychological phenomena such as norms of compliance, tax morale, or patriotism. An alternative set of explanations is that in reality, taxpayers may not believe the audit probability to be exogenous, or they may not know the actual audit probability — see Hashimzade et al. (2012) for a survey of the behavioral economics research applied to tax compliance. The latter case is relevant since in practice most taxpayers do not know the likelihood with which their tax return is audited by their country's tax authority. Most uneducated guesses are often an order of magnitude away from the actual audit rate. Relaxing the assumption of a known audit probability turns the taxpayer's compliance choice into a decision problem under ambiguity. In this framework, taxpayers do not know the true probability of audit, but may have prior beliefs about what probability is most likely. If they are pessimistic, they may assign a high likelihood to a very high audit rate, which might explain the high levels of compliance.

We report on a series of experiments testing the effect of norms of compliance on behavior by sampling our subject pool from three distinct populations: undergraduate students, who are the typical sample in economics experiments but have little to no experience with paying income tax; individuals in full-time employment who pay income tax through a third-party reporting system; and individuals who are self-employed and therefore self-report their income tax liabilities to the tax authority. We manipulate two standard policy levers in the classic models of tax compliance: the audit probability and the fine for non-compliance. We also consider the case where the audit probability is unknown.

The experiment was implemented on a sample of 520 individuals, of whom 200 were students, 200 were individuals who pay tax through a third-party system, and 120 self-employed taxpayers who file a return annually. We found very large subject pool differences, both in the level of compliance, as well as the responsiveness to changes in experimental parameters. Students were the least compliant subject pool, but also the most responsive to treatment changes, particularly to ambiguity in the audit probability, as well as changes in the fine for non-compliance. Self-employed taxpayers and taxpayers who pay through third-party reporting were more compliant and mostly non-responsive to different conditions.

A post-experimental survey uncovered that the vast majority of self-employed individuals may have exhibited high compliance levels in the experiment due to norms of honesty and compliance, in the sense that the experimental framing led them to translate their real-world behavior into the experimental task. To investigate the role of norms of tax compliance from outside the lab on behavior, we conducted an additional treatment in which any reference to tax, audits and fines was removed from the experimental materials. Average compliance in this treatment was reduced by half in the self-assessed sample, as well as the other two samples, highlighting the importance of norms in determining compliance in the lab.

The remainder of the paper is organized as follows. Section 2 contextualizes our work in the existing experimental literature on tax compliance experiments, both done in the lab and in the field. Section 3 outlines the hypotheses underpinning the experiment, and Section 4 describes the experiment. Section 5 presents the analysis and main results. Section 6 discusses the paper's results and Section 7 concludes the paper.

2. Previous tax compliance experiments

Our study contributes to a longstanding literature on tax compliance experiments. The earliest experimental study of tax compliance was conducted by Friedland et al. (1978) and since that study a steady flow of contributions have followed. The typical experiment takes a group of university student subjects who must choose how much of a given income to declare to the tax authority. The experimenter can vary the probability of audit, the tax rate or the fine for non-compliance. These variables can be known to the subjects with certainty, or they can be uncertain. The basic experimental design has not changed a great deal in the 30 plus years since the literature started — see Alm and McKee (1998) and Fonseca and Myles (2012) for reviews. The literature finds a small, but positive elasticity of tax compliance with respect to audit rates, and a smaller and also positive elasticity with respect to penalty rates.

The key advantage of laboratory experiments is that, unlike the field, the experimenter can accurately detect evasion, since income is perfectly observable in the lab. When conducting an empirical analysis on economics of crime, in whatever guise it may take, the econometrician is always impaired by the fact that she only works with data from those individuals who are caught. One never gets data on criminals who have never been caught, or those who cheated and then, for whatever reason, decided to stop. As such, we can never have measures of the deterrence aspect of fines, and only unreliable measures of the punitive effect.

There are two criticisms of laboratory experiments that, if taken at face value, limit the extent to which one can apply their findings to outside the lab. The first is the conceptual abstraction surrounding the task: there is typically little context surrounding the decisions subjects must make. While experiments are clearly good tests to the model (i.e. expected utility theory, or cumulative prospective theory) that underpins the theory of compliance, they may fail to be good tests of the

[1] There have been numerous extensions of this model, such as making labour supply endogenous, including a choice between employment in a formal and informal sectors, and increasing the complexity of the income tax (see the surveys of Pyle, 1991 and Sandmo, 2005) but the basic results are robust.

theory itself due to the lack of context inherent to the experimental task.[2] Experimentalists have attempted to circumvent this criticism by introducing features to the experimental task that make it closer to the behavior of interest in the real world. In the tax compliance case, experimental economists may introduce the need for subjects to earn their income, which they then declare for tax purposes. Bühren and Kundt (2014) conduct an experiment where they compare behavior in a tax evasion experiment in which income is either earned or randomly allocated. Their hypothesis is that, if their subjects have cumulative prospect theory preferences, earning their income shifts the reference point to the right, thus increasing the attractiveness of the evasion option. They implement a real-effort task in which they manipulate the difficulty level of earning income. Evasion rates are higher in the high difficulty treatment than in the control treatment where income is exogenous.

The second criticism is that the typical subject sample used in experiments may not be representative of the population. While some emphasize the role of financial incentives and argue that the validity of lab experiments is undiminished by the nature of the subject pool (Falk and Heckman, 2009), others claim that the putative control inherent to the lab may prove counterproductive if the task is inherently artificial to the subjects taking part in the experiment, and emphasize the importance of experience with the environment of interest in determining the external validity of any findings (Harrison and List, 2004).

In the context of tax evasion, the latter criticism equates to asking: why should one study tax evasion using a set of individuals who have never paid income tax? It is surprising that the experimental literature on tax evasion has only recently started to address this issue. Gërxhani and Schram (2006) experimentally studied compliance in two different countries (the Netherlands and Albania), and they looked at five separate subject pools: high school students, university students, high school teachers, non-academic university personnel, and university lecturers. The amount of under-reported income was higher in the Netherlands than in Albania, and higher for pupils and students than for teachers. Increasing the audit probability did not affect evasion in Albania, but did reduce evasion in the Netherlands. Alm et al. (2015) compared the behavior of undergraduate students to university staff and faculty, who pay their taxes through third-party reporting. They find students were less compliant than non-students, but had qualitatively similar responses to treatment effects. Bloomquist (2009) compares compliance behavior in the lab to behavior from random audits in the field and finds the two samples to be qualitatively similar.

In contrast to the experimental literature, there is a relative paucity of empirical work on tax compliance using field data. The emergence of randomized control trials (RCTs) in economics and their widespread use in policy has resulted in greater access to reliable data on tax evasion from the field. Slemrod et al. (2001) conducted an RCT on taxpayers in Minnesota. A letter was sent to a random subset of taxpayers who had filed a federal tax return during 1995, informing taxpayers that the return they would file that year would be “closely examined”. The data on the tax returns on the individuals receiving the letter were made available for the year of the intervention and the preceding year. The results showed that the effect of the letter depended on the level of income: low and middle income taxpayers who received the letter increased their reported income relative to the control group. The increase in reported income was also dependent on the source of income (higher among taxpayers declaring trade and business income than those declaring farm income), which indicated the effect of opportunity to evade. The surprise result was that the reported tax liability of the high income treatment group fell sharply relative to the control group. The authors proposed that this could be explained by the incentive to reduce the probability of an audit when the probability was less than one, as opposed to the belief that not all income would be discovered if audited for sure.

Kleven et al. (2011) report the results of an RCT in Denmark. The objective of the RCT was to ascertain the effectiveness of prior audits and different audit probabilities on reported income of individuals who pay their taxes either through a third-party reporting system or via self-reporting. The sample was 42,800 individuals in Denmark who were chosen to be representative of the population. In the initial year (2007) one half of the sample was randomly selected for rigorous audit treatment while the remainder were not audited. In the next year (2008) letters containing the threat of an audit were randomly sent to individuals in both groups. The individuals were not informed that they were part of an experiment. One group received a letter stating that an audit would certainly take place, a second group received a letter stating that half the group would be audited, and a third group received no letter. These different letters provided an exogenous variation in the probability of being audited. The effect of audits on future reported income was studied by comparing the audit and no-audit groups. This showed that audits had a strong positive impact on reported income in the following year. The effect of the probability of audit on reported income was analyzed using the threat-of-audit letter and no-letter groups. They found that evasion rates were close to zero among those who use third-party reporting, and significantly higher among those who self-report. Prior audits, and higher probability of future audit had a positive effect on compliance on self-assessed taxpayers but not on individuals who pay through third-party reporting. Also, the effects were stronger for the threat of an audit for certain than for the threat that half the group would be audited.

The main shortcoming of the RCT approach is that one cannot directly observe evasion, even if taxpayers are thoroughly audited. For example, cash transactions are, by their very nature, outside the scope of an audit, and unless a full audit of a company’s account is done — which is beyond the usual *modus operandi* of most tax agencies — the full extent of evasion can

[2] For a discussion of the artificiality critique of experimental economics, see Bardsley et al. (2010).

never be measured. One relies instead on variations in reported income as a proxy of compliance: if on average, reported income goes up as a result of a policy intervention, that must be due to higher compliance, rather than any other external factor. However, we cannot directly infer the impact of the policy on the fraction of taxpayers in full compliance, as well as the effect on the fraction of income reported by those who do not fully comply.

Our paper complements both literatures by setting up a real effort experiment, where there is individual level variation in income as well as accumulated wealth, but where evasion can be accurately detected. We are therefore able to estimate how the propensity to evade reacts to changes in income as well as accumulated wealth, as per Slemrod et al. (2001). We are also able to study the role of social norms of compliance by examining the behavior of different subject pools, in particular individuals who pay tax through third-party reporting and through self-assessment (like Kleven et al., 2011), as well as students who are the traditional subject pool in lab experiments.

3. Hypotheses

In the present section we present the hypotheses that guide our experiment. The Appendix contains the derivations of the results that form the basis of each hypothesis. The standard analysis of tax compliance follows the model of Allingham and Sandmo (1972) and Yitzhaki (1974). There, a single taxpayer must decide how much of her income to declare to the tax authority, conditional on a given (and known) tax rate, audit probability and non-compliance fine. A standard prediction in this set up is that increasing the audit probability should result in higher compliance. This is our first hypothesis.

Hypothesis 1. Increasing a known probability of audit will lead to higher levels of compliance.

In reality, taxpayers are not aware of the true probability of audit, and tax agencies do not usually publicize their auditing data to the general public. This means the taxpayer is making her evasion decision under ambiguity, rather than under risk. Under the model of ambiguity aversion proposed by Chateauneuf et al. (2007), ambiguity causes taxpayers to be responsive to the best and worst possible outcomes (i.e. being audited with probability zero or one, respectively). Existing survey and experimental evidence (Andreoni et al., 1998; Alm et al., 1992) suggests individuals may be ambiguity averse in the context of a tax compliance decision, and therefore assign a greater weight to the worst possible outcome (i.e. being audited). This constitutes the second hypothesis.

Hypothesis 2. Making the probability of audit unknown will lead to higher levels of compliance.

Our experimental design considers three distinct subject pools: students, taxpayers who pay taxes through third-party reporting, and self-assessed taxpayers. Individuals in each of the three subject pools have distinct experiences with the national tax authority, as well as norms of compliance, which likely emerge through professional social networks, as well as the history of audits. There are numerous ways in which social interaction can be introduced into the compliance decision. One way to do so is illustrated by the social custom model of Myles and Naylor (1996). This model assumes that there is a social custom that rewards compliance so that an honest taxpayer receives additional utility which is an increasing function of the proportion of taxpayers who do not evade. This captures the feature that evasion will cause more social prestige to be lost the more out of step the non-compliant taxpayer is with the remainder of society.

The expected utility model predicts that there should be no difference in the behavior of students and non-students in a compliance experiment. Social custom models argue otherwise: non-students will have been involved in the socialization process through which taxpayers absorb the social custom (of compliance). This opens the possibility that the different subject groups could have very different behaviors. In particular, Kleven et al. (2011) report lower compliance levels and higher responsiveness to changes in audit probabilities among taxpayers who self-report their income than those who pay through third-party reporting. This is consistent with the Myles and Naylor (1996) model, to the extent that norms of compliance should be stronger among taxpayers who pay taxes through third-party reporting than among self-assessed taxpayers. This leads to the next experimental hypotheses:

Hypothesis 3. Non-students will be more compliant and less responsive to changes in the probability of audit than students.

Hypothesis 4. Self-employed subjects will be less compliant and more responsive to changes in the probability of audit than subjects who pay their taxes through third-party reporting.

Finally, in all models of tax compliance, increasing the fine for non-compliance unambiguously leads to higher compliance, which is our last hypothesis.

Hypothesis 5. An increase in the fine for non-compliance will increase compliance.

4. Experimental design and procedures

In this section we will outline the experimental design, as well as the procedures we employed to recruit the experimental subjects, as well as the protocol used for running the experiments for each of the three different samples. For ease of

Table 1
Experimental design.

Fine level	Audit rate			
	P20N	P20	P40	UP
F100	30, 29, 27	35, 36, 31	35, 35, 30	30, 30, 32
F200		35, 35	35, 35	

Numbers are sample size for Student, PAYE and Self-Assessed subject pools.

exposition, we will refer to the undergraduate student sample as ‘students’, the third-party reporting taxpayer sample as ‘PAYE’, and the self-employed sample as ‘self-assessed’, to focus on their tax status, rather than their employment.[3]

4.1. The experiment

The experiment involved three main parts: the first part consisted of the main experiment testing tax compliance. The second part was a risk aversion elicitation experiment (Holt and Laury, 2002) and the third part consisted of a number of questionnaires eliciting personality measures, as well as attitudes towards tax paying. The four stages of the experiment included income generation, declaration of income for tax purposes, auditing and payoff generation. The experiment was run using z-Tree (Fischbacher, 2007). The currency in the experiment was the Experimental Currency Unit, ECU. A period in the experiment consisted of Stages 1–4 below, and there were 15 periods in total. The individual stages are:

Stage 1: Real-effort Task: Subjects performed a real-effort task for a fixed piece rate. This task was intended to induce a feeling of “ownership” of income. The task consisted of a set of 48 sliders on a screen (Gill and Prowse, 2012). Subjects earned a fixed payoff by placing each of the sliders at its halfway point.
Stage 2: Tax declaration: Subjects were told their Stage 1 income and had to declare their taxable income.
Stage 3: Auditing: Subjects were audited with a fixed probability p. If an audit occurred and subjects truthfully declared their taxable income, no penalty is levied. Otherwise, subjects paid the unpaid taxes in full plus a fine at rate f on unpaid tax.
Stage 4: Round payoffs: Subjects viewed their payoff for that round.

The three treatment variables are the probability of audit, the fine level, and the subject pool. We have four distinct probability conditions. Our baseline condition is P20, in which the probability of audit is equal to 20% and is public information. We study the effects of an increase in the audit probability using P40, in which the audit probability is equal to 40%. The UP treatment tackles the issue of ambiguity in the audit probability, by making it unknown to subjects, while being equal to 20% in practice. Finally, P20N is a payoff-equivalent version of P20, with a neutrally framed set up. We also consider two separate fine levels: in F100, non-compliant subjects who are audited pay a fine of 100% of unpaid tax; in F200, that fine is equal to 200% of unpaid tax.

Table 1 outlines the experimental design, as well as the number of subjects from each sample that participated in each treatment. The first number in each cell is the number of student subjects, the second number refers to the number of PAYE subjects and the third number is the number of self-assessed subjects. We only collected data on the student and PAYE samples in the F200 conditions, as the compliance level among self-assessed was close to 100% in the F100 conditions.

4.2. Subject recruitment

The subjects in our three pools were recruited using an email inviting them to participate in an experiment on decision making. This email made no explicit mention of the experiment being related to taxes, so as to avoid self-selection biases. Once subjects completed the experiment, they viewed a debrief form, which clarified the identity of the funding body, HMRC, but reiterated the fact that subjects’ and their decisions would remain strictly anonymous. Subjects were given the possibility to opt out of the study and have their data permanently deleted from the sample. Nobody exercised that option. Please see the Appendix for the recruitment and debrief materials.

4.2.1. The student sample

The student sample was recruited from a pool of voluntary undergraduate student subjects from the University of Exeter through the ORSEE system (Greiner, 2015). The age of our participants ranged between 18 and 49, with an average of 19.11; 64% of our subjects were male. We did not invite any one who had taken part in a similar experiment before. All the sessions took place in the experimental laboratory of the university. Upon arrival at the laboratory, subjects were assigned to their seat; once everyone was seated, no communication was allowed between subjects. The experimenters informed subjects they could not answer any questions from this point onwards. Subjects were paid individually in cash at the end of the

[3] PAYE stands for Pay-As-You-Earn, the system used in the UK for withholding income tax payments by companies on their employees’ behalf.

session. 15 ECU in the student sessions were worth £0.50. The average payment was £15.89, which included a show-up fee of £5.

4.2.2. The PAYE sample

The majority of the PAYE sample was recruited from a pool of voluntary subjects from across the UK run by a market research company, Saros Research. We also recruited PAYE taxpayers from businesses in the local area, as well as employees of the university. The age of our participants ranged between 19 and 66, with an average of 36.65; 59% of our PAYE subjects were male. Our lab and online PAYE samples did not differ in terms of average age (36.33 online vs 36.30 in the lab, $t = -0.01$, $p = 0.999$), but they did differ in terms of gender composition (65% males online vs 40% males in the lab ($\chi^2(1) = 8.90$, $p = 0.003$)). The subjects recruited by Saros Research are regularly paid to take part in focus group research and/or online surveys. To the best of our knowledge no subjects had taken part in economics experiments prior to our study taking place. Saros Research triaged subjects through an initial questionnaire which asked a battery of questions including a question asking whether they were full-time residents in the UK for tax purposes, and another question asking for their tax status. The subjects took part in the experiment from home or their place of work. To facilitate participation, we conducted sessions in the evenings between 6 pm and 9 pm. The experiment was run using z-Tree (Fischbacher, 2007). We provided subjects with software that connected their computer to our university servers. Subjects were asked to log on to the online system at a pre-designated time. PAYE subjects were paid through a bank transfer or through a cheque which was mailed to their home address.[4] The subset of PAYE subjects who resided in the university's area were recruited through ads and email. They travelled to the university laboratory to take part in the sessions, and they were paid in cash. 15 ECU in the PAYE sessions were worth £1.00. The average payment was £36.03 (average earnings by the lab and online samples were £40.01 and £34.74, which are statistically significant, $t = 6.48$, $p < 0.001$) which included a show-up fee of £20.

4.2.3. The self-assessed sample

The self-assessed sample was recruited from a pool of voluntary subjects from across the UK run by a market research company, ICM Research. The age of our participants ranged between 21 and 72, with an average age of 49.06; 78% of our subjects were male. Like the PAYE sample, these are regular paid subjects in market research who had never taken part in an economics experiment. The triage process was identical to that of the PAYE sample. Given the nature of the research and the subject pool, to minimize potential self-selection of subjects, as well as bias in choices in the experiment itself, ICM Research conducted all the recruitment and payment of subjects. Furthermore, we took extra measures to ensure anonymity of subjects, which were disclosed to subjects at the recruitment stage. Firstly, the researchers did not have any access to the names of subjects. Each subject was given a unique ID number, through which they would make their decisions. Only ICM Research could link names to ID numbers for payment purposes, but they could not access the experimental data itself.

To minimize direct contact with subjects, we designed a bespoke web-based software, which had the same visual interface as the software used by student and PAYE subjects. To access the experiment, subjects had to type their ID number plus a password. Subjects could log on at any time they wanted, within a week of receiving their log in information. However, once logged in to the experimental software, subjects had to complete the experiment within one hour of logging on. The experimental software did not allow subjects to log back in once an hour had elapsed. Subjects were paid by ICM via bank transfer. 15 ECU in the self-assessed sessions were worth £1.00. The average payment was £46.94, which included a show-up fee of £30.[5]

4.3. Experimental procedures

It is worth discussing the role of procedural differences in the data collection across subject pools. The primary difference is that the data from the student sample was collected in our laboratory, while most PAYE subjects and all self-assessed subjects took part in the experiment online — a subset of our PAYE sample undertook the experiment in our laboratory. While it is possible that this difference could lead to differences in behavior, we argue that the expected effect is *more* compliance by subjects in the lab. There is a much larger social distance (the degree of reciprocity that subjects believe exist in a social interaction) between the subjects who took part in the experiment online and the experimenters vis-à-vis lab participants and experimenters.[6] Online subjects are not being directly observed while making their decisions, and they do not meet the experimenter face-to-face when they collect their payoff, which is directly impacted by the decision to evade. As such, they do not risk any hypothetical shame from facing the experimenter after having broken the norm of honesty. Furthermore, in the case of our self-assessed sample, we employed a double-blind design, in which we did not know which subject took part in what treatment, and we used an intermediary to perform individual payments — something we made known to potential

[4] Upon signing up for the experiment, subjects provided us with their banking details through a secure web server, or with an address should they wish to be paid by cheque. Nobody declined to participate due to the method of payment.

[5] The different show-up fees reflected the different opportunity cost of time for each sample, compounded by the fact that non-student subjects did not typically take part in studies which demanded an hour-long commitment of their time outside working hours. For that reason, we also implemented a different exchange rate between ECU and pound sterling depending on the sample.

[6] Hoffman et al. (1996) study the impact of social distance on giving in dictator game and find higher social distance leads to more self-interested choices.

Table 2
Average compliance rate.

	Student				PAYE				Self-Assessed			
	P20N	P20	P40	UP	P20N	P20	P40	UP	P20N	P20	P40	UP
F100	0.30	0.61	0.64	0.73	0.51	0.84	0.80	0.84	0.44	0.93	0.93	0.93
	(0.23)	(0.35)	(0.36)	(0.30)	(0.34)	(0.24)	(0.23)	(0.23)	(0.30)	(0.19)	(0.15)	(0.16)
N	30	35	35	30	29	36	35	30	27	31	30	32
F200		0.79	0.84			0.81	0.89					
		(0.29)	(0.23)			(0.24)	(0.21)					
N		35	35			35	35					

Standard deviations in parentheses.

participants during the recruitment stage of the experiment. Therefore, if changing the mode of participation introduces any bias in compliance, it should be towards less compliance, due to the greater social distance.

Despite the differences in the recruitment of the three different subject pools, as well as the differences in the way they took part in the experiment itself (i.e. online vs the lab), the actual protocol of the experiment was the same across the three samples. Upon logging on to the software, subjects had 10 minutes in which to read the instructions on their computer screen, after which the experiment started. Subjects could not interrupt the experiment and log back on at a later time. Each period had a fixed duration; after that time elapsed, the next period commenced until the end of the experiment. Once all three parts of the experiment were complete, a debrief text appeared on the screen, which explained the purpose of the experiment, and were given the option to opt out of the study if they wished to do so. Subjects were paid after finished reading the debrief form. The experiment lasted for no longer than one hour. All recruitment materials and instruction sets are available in the Appendix.

5. Results

The analysis will focus on the subjects' compliance rate, which we define as the ratio of declared income to income earned in a given period of the experiment. This definition means that the compliance rate has a value between zero and one, which imposes some constraints on our method of data analysis when we go beyond analyzing average treatment effects. We elaborate on this issue in the appropriate sub-section.[7] We begin the analysis of results by looking at the effect of the different treatments on the average compliance levels. We then proceed to econometrically estimate the determinants of compliance.

5.1. Average compliance

Table 2 displays average compliance in the different treatments, for each of the three subject pools using the average behavior of each subject over the course of the experiment as the unit of observation.[8] We start by examining the effect of increasing the audit probability, when it is known (i.e. treatments *P*20 and *P*40). With the exception of treatment F200 in the PAYE sample, where we observe a significant difference ($z = 2.22$, $p = 0.026$), doubling the probability of audit has no statistically significant effect on average compliance, in all three subject pools. We now move to the effect of unknown audit rates on behavior. When we compare the treatment when audit rate is unknown (UP) to the equivalent treatment when the audit rate is known (P20), we do not observe a significant change in average compliance levels in students ($z = 1.55$, $p = 0.122$), PAYE ($z = 0.30$, $p = 0.766$) or self-assessed subjects ($z = 0.34$, $p = 0.731$).

Table 2 also reveals systematic differences in average compliance across the different subject pools. In the low fine treatments (F100), students exhibit lower average compliance than both PAYE (P20, $z = 3.21$, $p = 0.001$; P40, $z = 1.82$, $p = 0.069$; UP, $z = 1.54$, $p = 0.124$) and Self-Assessed taxpayers (P20, $z = 4.64$, $p < 0.001$; P40, $z = 3.97$, $p < 0.001$; UP, $z = 3.33$, $p < 0.001$). There are less pronounced but significant differences in average compliance between PAYE and self-assessed subjects. Surprisingly, the latter subject pool is more compliant than the former (P20, $z = 1.89$, $p = 0.059$; P40, $z = 2.60$, $p = 0.009$; UP, $z = 1.95$, $p = 0.050$).

We conclude our discussion of Table 2 by looking at the effect of increasing fine levels. Doubling the level of fine led to significantly higher average compliance levels in the student sample in both audit probability conditions (P20, $z = 2.33$, $p = 0.020$; P40, $z = 2.66$, $p = 0.008$), as well as the PAYE sample, but only marginally significantly so in the P40 condition (P20,

[7] About 4% of our data recorded subjects over-declaring their income. Unlike under-declarations, where it is impossible to distinguish between an individuals mistake and evasion, we can treat these observations as clearly errors and as such dropped those observations from the sample. While a frequent outcome is for a subject to make one mistake during the whole experiment, we found that 38% of over-declarations were made by 14 subjects (2.7% of the sample). We are confident that excluding these observations from the sample is simply ruling out the small subset of subjects who, despite our best efforts, perhaps did not understand the instructions quickly enough. Nevertheless our results would not qualitatively change if we had censored our dependent variable at 1.

[8] We use the non-parametric Mann–Whitney test, which we will denote as MW test, to compare mean compliance rates across samples and/or treatments throughout this sub-section.

$z = 1.16$, $p = 0.248$; P40, $z = 1.94$, $p = 0.052$). This meant that the differences in average compliance in the two subjects pools in the low fine conditions disappear in the high fine conditions (F200-P20, $z = 0.36$, $p = 0.359$; F200-P40, $z = 0.93$, $p = 0.178$).

5.2. Distribution of compliance

Restricting our analysis to treatment averages naturally ignores a great degree of heterogeneity in the data. Fig. 1 shows the set of histograms of compliance levels, using the average compliance by each subject as the unit of observation. While our measure of compliance is a continuous variable, Fig. 1 illustrates that when we aggregate individual compliance behavior over the course of the experiment, two behavioral types emerge: individuals who always declare their earnings in the experiment truthfully, and those who do not. Changing the probability of compliance from 20% to 40% led to no significant differences in the distribution of compliance in any of the three samples (students: $D = 0.114$, $p = 0.960$; PAYE: $D = 0.183$, $p = 0.498$; self-assessed: $D = 0.109$, $p = 0.988$, two-sample Kolmogorov–Smirnov (KS) test for equality of distributions). There was also no difference in the proportion of fully compliant subjects across the two conditions for any sample (students: $p = 1.000$; PAYE: $p = 0.341$; self-assessed: $p = 0.791$, Fisher's exact (FE) test).

Result 1. Doubling the audit probability results in no significant change in compliance in any of the three subject pools.

We now turn to the effect of ambiguity in the audit probability. We find introducing ambiguity in the audit probability leads to no significant change in either the distribution of compliance in all three subject pools (students: $D = 0.210$, $p = 0.383$; PAYE: $D = 0.106$, $p = 0.987$; self-assessed: $D = 0.153$, $p = 0.791$, KS test), or the probability of full compliance (students: $p = 0.381$; PAYE: $p = 0.464$; self-assessed: $p = 1.000$, Fisher's exact test, FE test).[9] This is our second result.

Result 2. Making the audit probability ambiguous did not result in any significant change in average compliance relative to the treatment in which the audit probability was known to be 20% for any of the three subject pools.

We now compare subject pools, keeping experimental parameters constant. We find that student subjects have a significantly lower proportion of fully compliant subjects, as well as a lower average compliance level by evaders relative to Self-Assessed taxpayers in all treatments (all comparisons significant at the 1% level).

When comparing students to PAYE taxpayers, with the exception of P20-F100, where there is a significant difference in the likelihood of full compliance ($p = 0.005$, FE test) and the distribution of compliance ($D = 0.354$, $p = 0.013$, KS test), there are no significant differences in either the likelihood of full compliance or distribution of compliance in all other treatments.

Result 3. Students exhibit lower average compliance than Self-Assessed taxpayers in all treatments and lower than PAYE taxpayers in the treatment with low fines and 20% audit probability. This difference is driven both by a lower likelihood of full compliance, and lower expected compliance levels by evaders.

We now compare the two non-student samples. We observe a marginally significant difference in the PAYE and self-assessed distributions of compliance in the P20 treatment ($D = 0.292$, $p = 0.076$, KS test), but no difference in the proportion of fully compliant subjects ($p = 0.214$, FE test). We do observe a significant difference between the two subject pools in the P40 treatment both in terms of the overall distribution ($D = 0.367$, $p = 0.014$, KS test) and the fraction of full compliers ($p = 0.048$, FE test). In the UP treatment, there is no difference in the overall compliance distribution ($D = 0.256$, $p = 0.189$, KS test), although there is a higher proportion of fully compliant subjects in the self-assessed sample.

Result 4. Self-assessed taxpayers exhibit higher average compliance than the PAYE sample.

We complete our analysis of aggregate data by looking at the effect of doubling the fine rate. In the student sample, doubling the fine rate when the audit probability is low leads to a significant change in the distribution of compliance ($D = 0.343$, $p = 0.019$, KS test), but it does not change the proportion of full compliers ($p = 0.265$, FE test), suggesting that the treatment effect operates solely among non-compliers. When the audit probability is high, doubling the fine leads to significantly higher proportion of full compliers ($p = 0.041$, FE test), as well as higher average compliance overall ($D = 0.314$, $p = 0.039$, KS test). In contrast, doubling the fine rate in the PAYE sample only had a marginally significant effect on the compliance distribution for the P40 condition ($D = 0.286$, $p = 0.076$). All other comparisons yielded non-significant differences.

Result 5. Doubling the fine rate for non-compliance resulted in significantly higher average compliance among students, but only had a marginally significant effect on behavior by PAYE taxpayers when the audit probability was high.

5.3. Dynamics of the individual compliance decision

We now turn to the dynamic effects of auditing on compliance. To this effect, we will take advantage of the fact that the experiment was repeated multiple times for each subject. As such, we can employ standard methods of panel data econometrics to investigate the role of audits on future compliance.

[9] The attentive reader may wonder why our model does not pick up an effect, despite the noticeable change in the histograms for the Student – UP and PAYE – UP conditions, namely an increase in the frequency of the top category. The reason for this slight increase is due to an increase in the fraction of subjects whose average contributions range between 0.96 and 0.99.

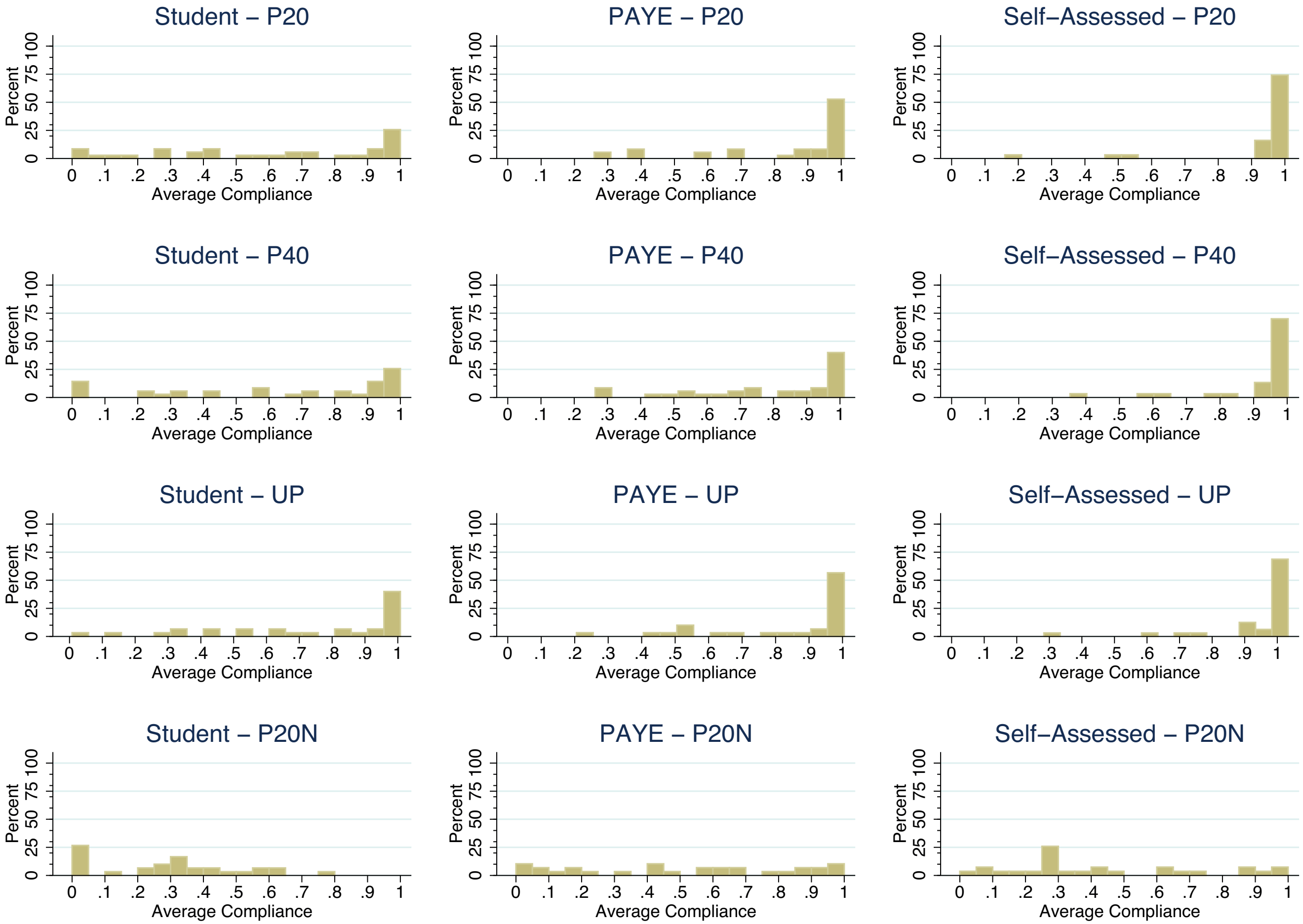

Fig. 1. Histograms of average proportion of declared income, F100 treatments.

Table 3
Random-effects Tobit estimates of determinants of compliance.

	(1)		(2)	
DV	$c_{it} \in [0, 1]$		$c_{it} \in [0, 1]$	
Constant	1.404***	(0.132)	1.340***	(0.151)
P20 × PAYE	0.727***	(0.079)	0.801***	(0.060)
P20 × Self-Assessed	1.120***	(0.127)	1.280***	(0.093)
P40 × Student	0.115**	(0.060)	0.168***	(0.062)
P40 × PAYE	0.439***	(0.063)	0.400***	(0.076)
P40 × Self-Assessed	1.074***	(0.092)	1.178***	(0.104)
UP × Student	0.198*	(0.113)	0.235**	(0.106)
UP × PAYE	0.612***	(0.071)	0.663**	(0.059)
UP × Self-Assessed	1.096***	(0.103)	1.238***	(0.086)
P20 × F200 × Student	0.473**	(0.069)	0.570***	(0.064)
P40 × F200 × Student	0.714***	(0.064)	0.898***	(0.052)
P20 × F200 × PAYE	0.375***	(0.073)	0.348***	(0.052)
P40 × F200 × PAYE	0.926***	(0.093)	1.009***	(0.081)
P20N × Student	−0.634***	(0.062)	−0.818***	(0.059)
P20N × PAYE	−0.276***	(0.071)	−0.444***	(0.076)
P20N × Self-Assessed	−0.320***	(0.077)	−0.501***	(0.073)
Income_{it}	0.010***	(0.002)	0.011***	(0.003)
$\text{Total Income}_{it-1}$	−0.001***	(0.0001)	−0.001***	(0.0001)
$(\text{Evade} \times \text{Not Audited})_{it-1}$	−0.337***	(0.033)		
$(\text{Evade} \times \text{Audited})_{it-1}$	−0.522***	(0.040)		
$(\text{Not Evade} \times \text{Audited})_{it-1}$	−0.226***	(0.052)		
$\text{Student} \times \text{Audited}_{it-1}$			−0.386***	(0.036)
$\text{PAYE} \times \text{Audited}_{it-1}$			−0.077**	(0.047)
$\text{Self-Assessed} \times \text{Audited}_{it-1}$			−0.015	(0.047)
Experience_i	0.011*	(0.002)	0.013***	(0.002)
Risk_i	−0.016	(0.006)	−0.017**	(0.006)
Male_i	−0.035	(0.028)	−0.030	(0.040)
Age_i	0.003	(0.002)	0.002	(0.002)
Extraversion_i	−0.014	(0.010)	−0.013	(0.012)
Agreeableness_i	−0.017***	(0.007)	−0.021***	(0.006)
$\text{Emotional Stability}_i$	−0.039***	(0.011)	−0.044***	(0.013)
$\text{Conscientiousness}_i$	−0.0001	(0.006)	−0.0003	(0.007)
Openness_i	−0.027***	(0.006)	−0.032***	(0.007)
N	5900		5900	
ρ	0.713	(0.024)	0.785	(0.017)

Bootstrapped standard errors in parenthesis.
* Significance at 10%.
** Significance at 5%.
*** Significance at 1%.

Given that the compliance decision is bounded between zero and one and we are likely to observe mass points at either end of the distribution, standard GLS is not appropriate. As such, we opted for a random effects two-limit Tobit model, in which:

$$c_{it} = x_{it}\beta + \nu_i + \varepsilon_{it} \tag{1}$$

where x_{it} is a vector of regressors for subject i in period t, β is the vector of coefficients to estimate, ν_i are i.i.d. $N(0, \sigma_\nu^2)$ and ε_{it} are i.i.d. $N(0, \sigma_\varepsilon^2)$ independently of ν_i. The observed data c_{it}^* is a potentially censored version of c_{it}. Our model will assume that $c_{it}^* = 0$ if $c_{it} < 0$, $c_{it}^* = c_{it}$ if $0 < c_{it} < 1$, and $c_{it}^* = 1$ if $c_{it} > 1$.

Table 3 summarizes the estimates from the random effects Tobit estimations. We consider two separate specifications, which we explain below. The coefficient on the treatment dummies in both random effects Tobit estimations reiterate the findings from the analysis of average compliance: expected compliance is not sensitive to either changes or to ambiguity in the audit rate. Students are the least compliant subject pool and self-assessed taxpayers are the most compliant. Finally, doubling the fine rate leads to significant changes in compliance. Importantly, we can now investigate the dynamic aspects of the compliance decision, namely the effect of past audits on present compliance, as well as the effect of individual heterogeneity, whether manifested through different ability, accumulated wealth in the experiment, risk attitudes, or personality traits.

We find a significant coefficient on Experience, which measures the number of years in the current occupation (students had a value of zero), which has a small positive coefficient. The coefficients on Emotional Stability, Agreeableness and Openness had negative and significant coefficients. This is consistent with the evidence from Alaheto (2003), who found in a survey of convicted felons that emotional stability and agreeableness were negatively correlated with the likelihood of committing white collar crime.

The real-effort nature of the experimental design allows us to exploit individual differences in ability and/or effort, which have a direct effect on the income each subject earned in a given period, as well as the accumulated income throughout the experiment.[10] On one hand, we find a positive and significant coefficient on Income_{it}, which we interpret as evidence that more productive subjects (either due to ability or higher effort) are less likely to evade.[11] On the other hand, the coefficient on $\text{Total Income}_{it-1}$ is negative and significant, which indicates a countervailing effect: the wealthier our subjects are, the more likely they are to evade.[12]

We now focus on the effect of audits on behavior. Regression (1) conditions the effect of an audit on whether a subject was fully complying or not, pooling across the three subject pools. To do so, we include three dummy variables interacting the decision of subject i in period t to evade or not (Evade_{it}; Not Evade_{it}) with the auditing outcome in period $t-1$ (Audited_{it-1}; Not Audited_{it-1}). The omitted category is the case where subject i did not evade in period t and was not audited in period $t-1$.

Starting with the case where the subject was not evading in period t, we observe a negative and highly significant coefficient on Not $\text{Evade}_{it} \times \text{Audited}_{it-1}$, indicating that expected compliance goes down in the period subsequent to an audit taking place. The same occurs in the case where subjects are evading: not only are the coefficients on both Evade Not Audited_{it-1} and Evade Audited_{it-1} negative and significant, as expected, but there is a significant difference between the two coefficients ($\chi^2(1) = 58.72$, $p < 0.001$), which indicates that even among evaders, the expected compliance goes down. Furthermore, the effect size of an audit on behavior by non-compliant subjects is economically similar, though significantly larger, to the effect of audits on compliant subjects ($\chi^2(1) = 41.19$, $p < 0.001$).

Having demonstrated that audits lead to lower compliance in the following auditing period, irrespective of whether one is a complier or evader, we wish to understand whether there are subject pool differences in the way compliance behavior changes following an audit. Regression (2) tackles this problem by replacing the aforementioned audit interaction dummy variables with a new set of interactions between $\text{Audited}_{i,t-1}$ and a dummy for each subject pool. We find dramatic subject pool differences: while the coefficients on all interaction dummies are negative, only those on the student and PAYE samples are significant. Furthermore, the coefficient on the student interaction dummy is significantly larger (in absolute terms) than that on the PAYE sample ($\chi^2(1) = 41.46$, $p < 0.001$). We find no statistically significant difference between the interaction coefficients on the PAYE and self-assessed samples ($\chi^2(1) = 1.31$, $p = 0.252$). In short, we find evidence for the bomb-crater effect (Mittone, 2006; Kastlunger et al. 2009) in our experiment, but that effect is driven primarily by the student sample. We find weaker evidence of that effect among the PAYE taxpayer sample and no evidence among Self-Assessed taxpayers. This constitutes our next result.

Result 6. Audits lead to a large fall in future compliance among students, as well as to a lesser extent, PAYE taxpayers. However, audits have no effect on future compliance behavior of Self-Assessed taxpayers.

The differences across samples in terms of compliance levels, as well as differences in changes with respect to changes in audit rates suggest that norms of compliance may have played a very important role in determining the compliance behavior of the self-assessed taxpayers.[13] In order to understand the role of norms, we conducted an extra treatment with the same parameter values as F100-P20, but in which the framing was neutral — we denote this treatment as P20N. That is, we removed all instances of "tax", "income declaration", "audit probability" and "fine" from the instructions and the text on the software interface. Subjects therefore faced the exact same decision problem, but without the normative context of compliance decision. We ran this treatment on the three subject pools.

Table 2 shows the average compliance in the neutral treatment was significantly lower in all three subject pools (students, $z = 3.59$, $p < 0.001$; PAYE, $z = 4.14$, $p < 0.001$; self-assessed, $z = 5.43$, $p < 0.001$, MW, test). Furthermore, while students remained the least compliant subject pool on average in P20N, the difference is no longer significant to that of self-assessed taxpayers (student vs PAYE: $z = 2.58$, $p = 0.0099$; Student vs Self-Assessed: $z = 1.47$, $p = 0.141$, MW test). We find no significant difference between average compliance of PAYE and Self-Assessed ($z = 0.632$, $p = 0.528$, MW test). Our analysis using the random effects Tobit estimations using period-level individual data reiterates our finding: compliance levels are lower in the P20N condition than in P20.[14]

[10] Gill and Prowse (2012) argue that this variable measures pure effort, rather than ability. However, their data was gathered from undergraduate students only, who grew up with computers, and therefore whose ability with a mouse is more or less homogeneous. It is plausible to argue that our older non-student subjects, who were not regular computer users from an early age, may have more heterogeneous ability levels. This is compounded by the fact that some non-student subjects may not have used a mouse.

[11] The reader will have noted from our description of the procedures, that the relative weight of the show-up fee on total payment is different between students and non-students. This is primarily due to the fact that students were more effective at solving the slider task, and therefore had more income to declare. Our Income variable controls for that discrepancy.

[12] This effect might be driven by perceived deservedness — see Bühren and Pleßner (2014).

[13] Evidence from a post experimental survey suggests that norms of compliance may have played an important role in the behavior by self-assessed taxpayers. Typical responses include: "In real life I would be too concerned that I would be caught if I cheated on my tax return. I reflected this attitude in the experiment."

[14] Model (1), P20N × Student = 0: $\chi^2(1) = 10.11$, $p = 0.002$; P20N × PAYE = P20 × PAYE: $\chi^2(1) = 21.50$, $p < 0.001$; P20N × Self-Assessed = P20 × Self-Assessed: $\chi^2(1) = 39.02$, $p < 0.001$. Model (2), P20N × Student = 0: $\chi^2(1) = 11.26$, $p < 0.001$; P20N × PAYE = P20 × PAYE: $\chi^2(1) = 23.40$, $p < 0.001$; P20N × Self-Assessed = P20 × Self-Assessed: $\chi^2(1) = 45.90$, $p < 0.001$.

In short, by removing the tax compliance context of the task, we turned a tax compliance decision into a risky choice under a loss frame: a truthful declaration of earned income equals a sure loss; the under-declaration of income is a gamble in which detection yields a bigger loss than under full compliance and non-detection gives zero losses. We observe evidence in line with the literature on loss aversion, with a larger proportion of income allocated to the risky prospect than to the sure loss. We can therefore separate the role of risk preferences from the role of norms, and we demonstrate the importance of the latter in the decision whether or not to comply. This is our final result.

Result 7. The high compliance level observed in our experiment, particularly in the Self-Assessed sample, can be attributed to norms of compliance.

6. Discussion

The main finding of the paper is the stark behavioral differences between students and non-students: the former are less compliant but more responsive to policy levers than the latter. This difference may be primarily driven by norms of compliance, insofar as our framing manipulation was able to reduce the difference in compliance levels between students and self-assessed to the point of not being statistically significant. It is possible that students are more loss averse than non-students, since we could not detect any role of risk aversion, at least insofar as our measure is able to do so. Dhami and al-Nowaihi (2007) show that cumulative prospect theory preferences can explain tax compliance levels with low audit probabilities. Self-assessed taxpayers and taxpayers who pay their taxes through third-party reporting seem to conform broadly to one of two types: compliant and non-compliant. The behavior of the former is consistent with the predictions by models of norm compliance rather than a risk-return trade-off.

We also find significant subject pool differences in their behavioral response to an audit. While students' average compliance falls following an audit — the 'bomb-crater' effect — we register a qualitatively similar, but much smaller effect among third-party reporting taxpayers. However we observe no such changes in behavior in the self-assessed subject pool. This finding is relevant when put into context of the lab and RCT literatures. On one hand, we demonstrate that this effect is broadly restricted to undergraduate students. On the other hand, we are able to reconcile the seemingly contradictory evidence from the two sets of randomized controlled trials on self-assessed taxpayers: Slemrod et al. (2001) report a sharp fall in reported tax liabilities after an audit among high-income groups after being exposed to the audit condition. In contrast, Kleven et al. (2011) report a *rise* in reported tax liabilities after an audit. The difference in these behaviors could be driven by taxpayers' beliefs about the actual audit rate, and whether that audit rate is exogenous. In our case where that audit rate is demonstrably exogenous, there is no subsequent change in compliance following an audit, since there is no reason why the compliance rate would change. This is unlike reality, in which taxpayers may have reason to believe the probability of being audited presently will strongly depend on their past behavior.

It is useful to compare our results to those of Alm et al. (2015), who study the behavior of undergraduate students to that of university staff in the United States. That study also found that students were less compliant than staff, but they responded similarly to treatment changes. However, the mean compliance rates in their study are much lower than in our case. While it is possible to comment on the differences in levels of compliance in both studies, we cannot say much about treatment effects, since the two papers study different questions. Alm et al. (2015) report on treatments which measure the impact of information on the filing procedure when the process itself is complex, as well as the role of tax deductions and tax credits. In that sense, the average compliance rate among students in baseline treatment in Alm et al. (2015), 0.618, is close to the average student compliance rate in our study, 0.61. We observe a larger difference in the compliance rate among non-students: 0.84–0.93 in our non-student samples compared to 0.795 in Alm et al. (2015). It is also worth pointing out that the subject pool differences between the two studies are not restricted to a difference in nationality: while most employees in the UK will not file a tax form at the end of the fiscal year, all taxpayers in the US must do so. In that sense, the sample of workers in Alm et al. (2015) is a hybrid of our PAYE and Self-Assessed samples. Replicating our design in the US, or the design by Alm et al. (2015) in the UK would be one way to verify whether cross-country differences are the reason behind this discrepancy. We leave this for future research.

7. Conclusion

Our experiment bridges two strands of the empirical literature on tax compliance: the longstanding literature on laboratory experiments and the nascent literature using randomized controlled trials. The former has relied primarily on undergraduate students as subjects and in most cases, taxable income was exogenously allocated, while in the latter the primary subject pool are actual taxpayers, whose taxable income is earned. We do so by conducting a laboratory experiment with three types of subjects who have very different experiences in the way they pay tax. We study the behavior of students, who have no tax experience; workers who pay their income tax through third-party reporting systems; and self-employed individuals who file their own tax returns.

We find stark behavioral differences between students and non-students. The former are less compliant but more responsive to policy levers than the latter. Through a framing manipulation, we show that behavior in the three samples is driven by social norms of compliance, with the strongest effect coming from the Self-Assessed sample. A follow-up survey on the self-assessed sample corroborates the experimental findings. Our findings raise the obvious question of whether students are appropriate subjects for tax compliance experiments. Based on our evidence, undergraduate students may be more appropriate to study questions which require highly motivated, profit-maximizing agents. Their behavior will likely provide an upper bound for the effectiveness of a particular policy tool that relies on financial penalties. Self-assessed taxpayers may be more appropriate to study questions pertaining to norms of compliance, and how such norms propagate through social and professional networks.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jebo.2015.09.015.

References

Alaheto, T., 2003. Economic crime: does personality matter? Int. J. Offender Ther. Comp. Criminol. 47, 335–355.
Allingham, M.G., Sandmo, A., 1972. Income tax evasion: a theoretical analysis. J. Public Econ. 1, 323–338.
Alm, J., Bloomquist, K.M., McKee, M., 2015. On the external validity of laboratory policy experiments. Econ. Inq. 53 (2), 1170–1186.
Alm, J., McClelland, G.H., Schultz, W.D., 1992. Why do people pay taxes? J. Public Econ. 48, 21–38.
Alm, J., McKee, M., 1998. Extending the lessons of laboratory experiments on tax compliance to managerial and decision economics. Manag. Decis. Econ. 19, 259–275.
Andreoni, J., Erard, B., Feinstein, J., 1998. Tax compliance. J. Econ. Lit. 36, 818–860.
Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., Sudden, R., 2010. Experimental Economics: Rethinking the Rules. Princeton University Press, Princeton/Oxford.
Bloomquist, K., 2009. A comparative analysis of reporting compliance behavior in laboratory experiments and random tax audits. Natl. Tax Assoc. Proc. 2009, 113–122.
Bühren, C., Kundt, T.C., 2014. Does the level of work effort influence tax evasion? Experimental evidence. Rev. Econ. 65 (2), 137–158.
Bühren, C., Pleßner, M., 2014. The trophy effect. J. Behav. Decis. Mak. 27 (4), 363–377.
Chateauneuf, A., Eichberger, J., Grant, S., 2007. Choice under uncertainty with the best and worst in mind: neo-additive capacities. J. Econ. Theory 137, 538–567.
Dhami, S., al-Nowaihi, A., 2007. Why do people pay taxes? An explanation based on loss aversion and overweighting of low probabilities. J. Econ. Behav. Organ. 64, 171–192.
Falk, A., Heckman, J., 2009. Laboratory experiments are a major source of knowledge in the social sciences. Science 326, 535–538.
Fischbacher, U., 2007. z-Tree – Zurich toolbox for readymade economic experiments. Exp. Econ. 10, 171–178.
Fonseca, M.A., Myles, G.D., 2012. A survey of experiments on tax compliance. HMRC Research Report 198.
Friedland, N., Maital, S., Rutenberg, A., 1978. A simulation study of income taxation. J. Public Econ. 10, 107–116.
Gërxhani, K., Schram, A., 2006. Tax evasion and income source: a comparative experimental study. J. Econ. Psychol. 27, 402–422.
Gill, D., Prowse, V., 2012. A structural analysis of disappointment aversion in a real effort competition. Am. Econ. Rev. 102, 469–503.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1 (1), 114–125.
Harrison, G.W., List, J.A., 2004. Field experiments. J. Econ. Lit. 42, 1009–1055.
Hashimzade, N., Myles, G.D., Tran-Nam, B., 2012. Applications of behavioural economics to tax evasion. J. Econ. Surv. 27 (5), 941–977.
HM Revenue and Customs, 2013. Measuring tax gaps: 2013 edition. HMRC Corporate Communications.
Hoffman, E., McCabe, K., Smith, V., 1996. Social distance and other-regarding behavior in dictator games. Am. Econ. Rev. 86, 653–660.
Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. Am. Econ. Rev. 92, 1644–1655.
Inland Revenue Service, 2012. IRS releases new tax gap estimates; compliance rates remain statistically unchanged from previous study, Available at: http://www.irs.gov/uac/IRS-Releases-New-Tax-Gap-Estimates;-Compliance-Rates-Remain-Statistically-Unchanged-From-Previous-Study.
Kastlunger, B., Kirchler, E., Mittone, L., Pitters, J., 2009. Sequences of audits, tax compliance, and taxpaying strategies. J. Econ. Psychol. 30, 405–418.
Kleven, H.J., Knudsen, M.B., Kreiner, K.T., Pedersen, S., Saez, E., 2011. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. Econometrica 79, 651–692.
Mittone, L., 2006. Dynamic behaviour in tax evasion: an experimental approach. J. Socio-Econ. 35, 813–835.
Myles, G.D., Naylor, R.A., 1996. A model of tax evasion with group conformity and social customs. Eur. J. Polit. Econ. 12, 49–66.
Pyle, D.J., 1991. The economics of taxpayer compliance. J. Econ. Surv. 5, 163–198.
Sandmo, A., 2005. The theory of tax evasion: a retrospective view. Natl. Tax J. 58, 643–663.
Slemrod, J., Blumenthal, M., Christian, C., 2001. Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota. J. Public Econ. 79, 455–483.
Yitzhaki, S., 1974. A note on income tax evasion: a theoretical analysis. J. Public Econ. 3, 201–202.