When Lab Subjects Meet Real People: Comparing Different Modes of Experiments*

Pablo Beramendi

Duke University

Raymond M. Duch and Akitaka Matsuo

Centre for Experimental Social Sciences
Nuffield College
University of Oxford

December 25, 2014

*Paper prepared for the *2015 Asian Political Methodology Conference*, Academia Sinica, Taipei, Taiwan **Preliminary Work**

1 Introduction: Lab and Online Experiments

This research note is a first attempt to compare laboratory and online crowd-sourced subjects participating in an identical experiment involving strategic interactions. It builds on the tax compliance experiments conducted by Beramendi and Duch (2014). We compare the characteristics of the two different subject pools; their behaviour in simple non-interactive experimental games; and finally the choices they make in the tax compliance experiment that involve strategic interactions. Of particular interest is the latter question: Whether the choices of subjects in a lab, where they interact with subjects in close proximity, resemble those of subjects online where they "interact" with subjects who clearly not proximate (nor is the interaction in real time).

Conducting experiments in an experimental lab setting has a number of advantages, particularly when subjects interact with each other in real time (Morton and Williams 2009). And this is the case for public goods games in which subjects contribute to a common pool and are compensated depending on the size of these contributions. Lab experiments are the gold standard for ensuring internal validity. If we think that treatment effects are likely to interact with characteristics of convenience samples, and we are concerned with external validity, then a more diverse subject pool should be considered. Belot, Duch and Miller (2009) have compared the traditional student convenience samples that make up the lab subject pool with non-student subjects and find that certainly with respect to games involving other-regarding preferences students can differ significantly from non-students. The tax compliance experiments are a good case in point because they were designed to provide insights into redistributive preferences and we strongly suspect that the population is extremely heterogeneous with respect to these preferences.

For experiments of this nature the experimental lab represents a critical phase of the research design. It is in the lab where we can ensure the internal validity of the design. The tax compliance experiments were initially designed and implemented in the Nuffield

Centre for Experimental Social Sciences lab. Results from four sessions with a total of 66 subjects were the basis for assessing the internal validity of the experimental design and whether the treatment effects were informative.

The lab phase is also the foundation for the key innovation that we implement as part of the online version of the experiment. Subjects in the lab and online versions of the tax compliance experiments essentially play the same game and make choices in the identical contexts. In the lab version of the experiment subjects are randomly assigned to groups of six. And the collective decisions of the six subjects affect the payoffs of each subject. The real challenge for the online version of the tax experiment is matching the online participant to a similar group of subjects. It is difficult to do this matching in real time. As we will explain in much greater detail below, online subjects are assigned, through a substitution strategy, to the lab groups that have already played the game. Payoffs to the online subjects are then determined by their choices and the choices that had already been made in the lab experiments.

We replicate the lab experiment with online subjects because we are concerned with external validity. The subject pool employed for the online tax compliance experiment is Amazon Mechanical Turk (MTurk). Crowd-sourced subject pools are not representative of specific populations but there is considerable evidence that they are a reasonable basis for estimating heterogeneous effects. Experimenters frequently make use of subjects recruited from crowdsourcing platforms or other convenience online subject pools because they are convenient and cost efficient. MTurk is one of the most frequently used online crowd-sourcing platform for micro-tasks, called Human Intelligent Tasks (HITs). In this labor market, requesters who need workers for a relatively simple but labor intensive tasks can advertise the HITs with a minimal cost as low as 1 cent per HIT. People who have already registered as a workforce at MTurk (called workers) browse the list of HITs, and can accept and complete HITs. MTurk becomes a popular platform in social sciences

because of the cost efficiency compared to the traditional method for data collection in particular for survey or experimental research. Experimental research employing MTurk subjects has been widely published in leading journals in political science (e.g. Healy and Lenz 2014, Malhotra and Margalit 2014, Grose, Malhotra and Parks Van Houweling 2014), economics (e.g. Horton, Rand and Zeckhauser 2011, Olea and Strzalecki 2014, **?**) and sociology (e.g. Tsvetkova and Macy 2014).

Researchers have explored whether the research findings in survey research using traditional sampling both online and face-to-face can be replicated using the subjects from MTurk. Berinsky, Huber and Lenz (2012) reports that many of the established results from survey experiments using framing as a treatment can be replicated employing the MTurk subject pool (see also Crump, McDonnell and Gureckis 2013, Rand 2012). Other studies have claimed that MTurk subjects are even preferable than traditional samples. For instance, **?** show that the attrition rate for a multiple wave survey is lower for MTurk samples than some other experiments; they also contend that the MTurk sample is no less representative than conventional telephone surveys of the U.S. population.[1] Also, Grose, Malhotra and Parks Van Houweling (2014) maintain that MTurk subjects are more suitable for some types of treatments, such as reading lengthy texts, than online national representative samples because "MTurk respondents are 'workers' who are accustomed to completing lengthy tasks in exchange for payment."

There are some issues associated with MTurk subjects, and researchers need to use the best practice for implementing the research based on the MTurk subjects (Paolacci, Chandler and Ipeirotis 2010, Horton, Rand and Zeckhauser 2011). For example, they should screen out the subjects who do not pay enough attention to survey protocols and instructions. Berinsky, Margolis and Sances (2014) suggest implementing multiple screener questions in an unobtrusive fashion.

This essay is primarily concerned with whether there are mode effects associated with

[1] Their comparison is to an unweighted CBS survey.

the implementation of the tax compliance experiment with an online subject pool. Since we have implemented essentially identical experiments both in the lab and online we are able to address this question. And in particular we are interested in mode affects that are associated with experiments involving strategic interaction amongst subjects. As we point out above there have been some comparisons between experiments conducted with online subject pools and other types of subject pools. These studies though have not focused on experiments that involve strategic interaction amongst individuals.

There are experiments in which online subjects playing real-time synchronic games (e.g. Mason and Suri 2012), but most studies in behavioural economics using crowd-sourced subjects do not use realtime interaction. Many studies use either one-shot game without feedback (e.g. Horton, Rand and Zeckhauser 2011) or sequential interaction in which subjects leave a virtual lab and come back after several days (e.g. Tsvetkova and Macy 2014).

In our online experiment we also did not implemented a real-time synchronic interaction. Rather, what we do is to put online subjects in a virtual laboratory using the data obtained from the laboratory sessions held prior to the online session. When online subjects come in the virtual laboratory, they are randomly assigned to one of the groups in the laboratory sessions, and based on their responses they are substituted with one of the laboratory subjects and participate in the tax compliance experiment as a surrogate.

2 Experimental Setting

The tax compliance experiment was designed and implemented by Beramendi and Duch (2014). Their concern is understanding compliance with redistributive taxation given that such compliance is the foundation for the modern welfare state. They argue that beyond arguments about the monitoring ability and the efficiency of state institutions and normative concerns for the worse-off in society, the distributive incidence of

the public sector over the long run drives the level of compliance. In other words, the distribution of net benefits (as determined by the progressivity of taxes and transfers) is a key determinant of what individuals perceive as collectively fair, and therefore drives their willingness to contribute to the common pool. They test the argument with a series of lab and online experiments in which subjects are randomly assigned to different fiscal treatments, treatments that approximate the variation in the design of national fiscal systems around the world.

This essay describes a novel strategy for replicating an interactive lab experiment in an online environment. Of particular concern is whether subjects in the online version of the lab experiment behaved similarly to the subjects in the lab. The tax compliance experiment has two behavioural phases that are the basis for our comparisons: 1) subjects in both modes do real effort tasks; 2) both sets of subjects are required to report their earned income under the pressure of being audited and free from being audited (allowing them to cheat without penalties). In addition, subjects perform other tasks that again are the basis for comparison across the two modes: 1) a Dictator Game to measure other-regarding preferences ; 2) a Lottery game to measure risk preferences; and 3) and a integrity test.

2.1 Laboratory Experiment

The lab experiments were conducted at the CESS experimental lab with 66 University of Oxford students from the CESS subject pool. Participants receive printed instructions at the beginning of each module, and instructions are read and explained aloud. The experimental sessions were conducted from November 22 to December 3, 2013. In some modules of the experiment we offered the subject earnings in Experimental Currency Unit (ECU). The conversion rate is 1000 ECUs to 1 British Pound.

The tax treatments consist of ten rounds each. Table 1 summarises the treatments

that are implemented in these two modules of the experiment. Prior to the tax treatments, participants are randomly assigned to groups of six and we follow a partner matching. Thus, the composition of each group remains unchanged for the two tax treatment modules – in the first tax treatment module, subjects face an audit rate of 0% and in the second the audit rate is set at 10%. Each round of these two tax modules is divided in two stages. In the first stage subjects perform a real effort task. This task consist of computing a series of additions in one minute. Their Preliminary Gains depend on how many correct answers they provide, getting 150 ECUs for each correct answer.

We conduct a total of four different sessions that are summarised in Table 1. Note that in each session the tax rate is consistent – what varies across tax modules is the audit rate. Once subjects have received information concerning their Preliminary Gains, participants are asked to declare these gains. A certain percentage or "tax" (that depends on the fiscal regime treatment) of these Declared Gains is then deducted from their Preliminary Gains. In sessions 1 and 3, the group revenues are distributed equally amongst the six participants – these reflect fiscal regimes in which the distribution of social benefits are non-progressive. In sessions 2 and 4, group benefits are distributed in a progressive fashion such that the poorest two group members receive 50 percent of the group revenues; the middle income subjects receive 35 percent; and the richest two are given only 15 percent.

Treatments: Audit Rates. There were two audit rate treatments in the experiments: 0% and 10%. In the former treatment the subjects' Declared Gains are not subject to verification; in the latter 10% audit treatment, subjects have a 10 percent probability of being subjected to an audit. Subjects were randomly selected to be audited. When audited the Declared Gains are compared with the actual Preliminary Gains in order to verify these two amounts correspond. If the audit finds a discrepancy between the Preliminary and Declared gains an extra amount is deducted from the Preliminary Gains. The extra amount corresponds to 50% of the observed discrepancy. In addition, the

regular deduction applies to the Preliminary Gains and not to the declared amount. Deductions applying to the four group members are then pooled and distributed amongst those members in accordance with the redistribution rate in Table 1.

Table 1: Summary of Tax Compliance Experimental Treatments

Session	Subjects	Number of Groups	Tax Terciles	Benefits
1	24	4	1) 28% 2) 30%; 3) 32%	1) Rev/6; 2) Rev/6; 3) Rev/6
2	18	3	1) 20% 2) 30%; 3) 43%	1) 50% Rev; 2) 35% Rev; 3) 15% Rev
3	12	2	1) 42% 2) 45%; 3) 47%	1) Rev/6; 2) Rev/6; 3) Rev/6
4	12	2	1) 42% 2) 45%; 3) 47%	1) 50% Rev; 2) 35% Rev; 3) 15% Rev
Total	66	11		

At the end of each round participants are informed of their Preliminary and Declared gains; whether these two amounts have been audited; the amount they receive from the deductions in their group; and the earnings in the round. At the end of each tax module one round is chosen at random, and their earnings are based on their profit for that round. Participants are only informed of their earnings for each tax module at the end of the experiment.

Heterogeneity. The experiment also measured a set of auxiliary variables designed to allow us to explore heterogeneity in treatment effects. The demographic variables were gender and income. We included a measure of trust and a measure of ideological self-identification. And finally we measured other-regarding preferences with a version of the Dictator Game and measured risk aversion with a standard Holt-Lowry game.

Dictator Game. In order to evaluate arguments regarding other-regarding preferences and attitudes about redistributive taxation we included in the first module a Dictator Game. Subjects are asked to allocate an endowment of 1000 ECUs between them and another randomly selected participant in the room. Participants are informed that only half of them will receive the endowment, and the ones who receive the endowment will be randomly paired with those who don't. However, before the endowments are distributed and the pairing takes place, they may allocate the endowment between themselves and the other person as they wish if they were to receive the endowment.

Risk Aversion. Concern about job or status security is hypothesized to shape redistribution preferences. Risk averse subjects should be most enthusiastic about redistributive taxation. The fourth and last module of the experiment consists of a lottery-choice test consisting of ten pairs, which is based in the low-payoff treatment studied in (Holt and Laury 2002). The lottery choices (shown in Table 2) are structured so that the crossover

point to the high-risk lottery can be used to infer the degree of risk aversion. Subjects indicate their preferences, choosing Option A or Option B, for each of the ten paired lottery choices, and they know one of these choices would be selected at random ex post and played to determine the earnings for the option selected.

Table 2: Lottery Choices

Lottery	Option A	Option B
1	10% of 2.00£, 90% of 1.60£	10% of 3.85£, Bs. 90% of 0.10£
2	20% of 2.00£, 80% of 1.60£	20% of 3.85£, Bs. 80% of 0.10£
3	30% of 2.00£, 70% of 1.60£	30% of 3.85£, Bs. 70% of 0.10£
4	40% of 2.00£, 60% of 1.60£	40% of 3.85£, Bs. 60% of 0.10£
5	50% of 2.00£, 50% of 1.60£	50% of 3.85£, Bs. 50% of 0.10£
6	60% of 2.00£, 40% of 1.60£	60% of 3.85£, Bs. 40% of 0.10£
7	70% of 2.00£, 30% of 1.60£	70% of 3.85£, Bs. 30% of 0.10£
8	80% of 2.00£, 20% of 1.60£	80% of 3.85£, Bs. 20% of 0.10£
9	90% of 2.00£, 10% of 1.60£	90% of 3.85£, Bs. 10% of 0.10£
10	100% of 2.00£, 0% of 1.60£	100% of 3.85£, Bs. 0% of 0.10£

Integrity Test. Subjects who think that dishonest behaviour can be justified depending on a context would also consider that evading tax can be justified in the situation where the tax scheme is unfair to a social group they are affiliated. To explore the possibility of such explanation, we also measure an integrity score of respondents using a test developed by the Essex Centre for the Study of Integrity. In this test, respondents are asked to rate a range of activities can ever be justified (Table 3) . The respondents rate each item using a four point scale from 1) Never justified to 4) Always justified. In the aggregate index is the summation of this score which ranges from 10, highest integrity, to 40, lowest integrity.

Table 3: Integrity Test

Content of Item
A. Avoiding paying the fare on public transport.
B. Cheating on taxes if you have a chance.
C. Driving faster than the speed limit.
D. Keeping money you found in the street.
E. Lying in your own interests.
F. Not reporting accidental damage you have done to a parked car.
G. Throwing away litter in a public place.
H. Driving under the influence of alcohol.
I. Making up things on a job application.
J. Buying something you know is stolen.

2.2 Online Experiment

The Beramendi and Duch (2014) experimental research design called for implementing an identical version of the lab experiment with a much larger and more diverse online subject pool. As the previous section makes clear the lab experiment subjects are assigned to groups and their final payments are conditional on the behaviour of the other subjects in the group. Replicating this group assignment and real-time feedback in an online setting is challenging particularly for a relatively large number of subjects. We solve this challenge by matching online subjects to the groups that were formed as part of the lab experiment.

The sequence of the modules in the experiment is exactly the same as the lab experiment. The only difference is the number of rounds for the tax compliance game. To prevent the subjects from losing attention to the experiment, we reduce the number of round to the half. In the lab experiment, subjects play in total of 20 rounds of the tax compliance game (10 with auditing, 10 without auditing), while in the online version of this experiment, subjects play 10 rounds (5 with auditing, 5 without auditing). One of the ten rounds is randomly selected for the payment.

Another difference between the lab and online version of this experiment is the currency conversion between the British Pound Sterling for lab to the US Dollar for online. We simply exchange a pound to a dollar: In the online experiment, a thousand ECU is equal to a dollar, and for the risk elicitation, the currency units are changed to dollars.

The wording used in the questions is exactly the same as the lab experiment. However, we trimmed down the instructions given to the subjects. In the laboratory experiments, the experimenter reads the instruction aloud, and experimental subjects cannot skip the instruction to proceed to the next screen until the experimenter finishes the recitation. Since this is not the case for online experiments, we have tried to make sure that the experimental subjects receive the instructions by making the instructions shorter and clearer as well as by setting a minimum time to spend in one page of instruction before proceeding.

We recruited 500 subjects from MTurk. The worker qualifications we used were country of residence (United States), number of completed HITs (more than 500), and rate of previous HITs approved by requesters (99 percent or higher). We published the HIT on October 3, 2014 and obtained the 500 completion within a day. When MTurk workers accepted the HIT for this experiment, they were redirected to a virtual laboratory hosted on Qualtrics, an online survey platform, and randomly assigned to one of the 11 laboratory groups (see Table 1). In the tax compliance game, online participants played exactly the same real effort tasks, which are two-digits number additions for 30 seconds. Each correct answer is converted to 150 ECUs.

A critical component of the experimental design is assigning subjects to a tax bracket depending on how they perform in the real effort task compared to other subjects in their group. For the online experiment we match each subject to a subject from the previous lab experiment – the matching is based on the similarity of their Preliminary Gains. The online subject's tax bracket and after-tax revenues are determined by the subject and

group to which they are matched. This matching is accomplished by linking a database on our web server to Qualtrics. A survey experiment using the similar technology is found in Boas and Hidalgo (2013). We set up an online database that stores the laboratory data regarding the experimental parameters (such as the audit rate and tax scheme) as well as the real effort task performance and tax compliance in the laboratory sessions. When an online subject is assigned to one of the lab groups, Qualtrics software accesses the database through a web-service function and retrieves the information about the tax scheme for the group. The information will be shown on the subject's instruction screen.

Qualtrics' access to the laboratory data works in the following manner: In each round of the online tax compliance game, the Qualtrics survey program records the Preliminary Gain of an online subject, and then interacts with a program placed on our web server. The Qualtrics survey program carries information about the group affliction and Preliminary Gain of an online subject. The program on our server receives the information and then retrieves the lab data of the group for the round in order to compare the Preliminary Gain of the online subject with the Preliminary Gains of the lab subjects. The server program determines a lab subject whose Preliminary Gain is the most approximate to the online subject, and this lab subject is replaced with the online subject for the round. After determining the lab subject to be replaced, the server side program retrieves the data on total deductions of other five subjects, and then passes the data on the total contributions, tax bracket of online subjects, and audit information on Qualtrics. These information is shown on the experimental screen before the online subject is prompted to report their income. The online subject's total earnings for this round are then calculated as follows: the deduction to the Preliminary Gain of online subject's is made at an appropriate rate (including penalty if audited), the deduction is added to the common pool of the tax revenue from other five players; and the online subject's share of redistribution of the tax revenue is then calculated. At the end of each round online participants receives

a feedback regarding the round's results, which is equivalent to the information the lab subjects have received.

3 Lab versus Online Experiments Compared

The Beramendi and Duch (2014) decision to implement the online versions of the tax compliance experiment was in part a concern with heterogeneous treatment effects. Significant heterogeneity in treatment effects might raise questions about the treatment effects measured in the lab. The tax compliance experiments included a number of behavioural measures that are plausibly correlates of tax compliance. Evidence that these measures significantly differ between lab versus online experiments might be cause for questioning the external validity of the lab results. We begin with a comparison of these items.

A second set of comparative analyses will focus on behaviour specifically related to the tax compliance treatments. There are two key behavioural features of the tax compliance experiments. First, subjects undertake real effort tasks that determine their earnings. We have subjects undertake real effort tasks for which they earn money because we expect that their responses to taxes on earned income are more informative than their reactions to taxes on windfall endowments received from the experimenter. Our expectation is that both lab and online subjects would treat the real effort tasks with similar levels of seriousness and they would register similar outcomes. Both subject pools are incentivised although the online subjects earn less and for many this is one of a large number of HITs that they perform. And as we pointed out earlier there is some question about how attentive crowd-sourced subjects are to these such tasks. Significant differences would raise question about the employment of real effort tasks with online experimental subjects.

A core feature of the Beramendi and Duch (2014) tax compliance experiment is "cheating". The design is intended to facilitate non-compliance with tax policies. As we pointed

out earlier, the basic design is similar for both online and lab versions of the experiment. But there are factors that might result in differences across the subject pools. Obviously the subject pools differ: One might imagine, for example, that subjects in a lab experiment might be more hesitant about cheating given that they actually physically interact with other participants in the lab. There is little questioning the anonymity of online subjects and this might exaggerate cheating levels.

An additional factor that might affect the choices subjects make relates to the design innovations that were necessary in order to replicate the tax compliance experiments online. The innovation here is assigning online subjects to groups of lab subjects who had already played the game – in fact, each online subject was substituted for a real lab subject who had already played the game. There is no deception in how we implement the assignment of online subjects – they are told that they are being paired with subjects who had already made their decisions about how much of their income to declare. Our expectation is that this relatively minor change in the tax compliance experiment instructions should not result in the online subjects complying with tax rates differently than lab subjects. One of the goals of the comparisons we implement is to test this conjecture.

3.1 Difference in the demography and behavioural measures

Basic Demographics. Lab experiments typically do not elicit extensive demographic information about subjects. Therefore we restrict our comparisons to the age and gender of subjects. Figure 1 indicates that the gender distribution of subjects in the lab and online are quite similar. Figure 2 confirms what we might expect; there are age differences in the two subject pools. We know that MTurk workers tend to be younger than population survey samples (Berinsky, Huber and Lenz 2012), but since the laboratory subjects in this study are undergraduate students, the laboratory subjects are even younger on average.

Figure 1: Gender of Subjects

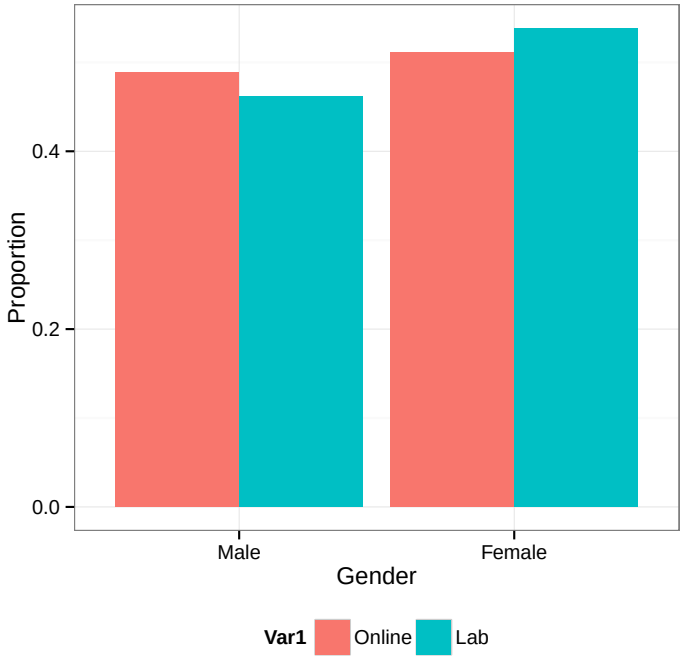
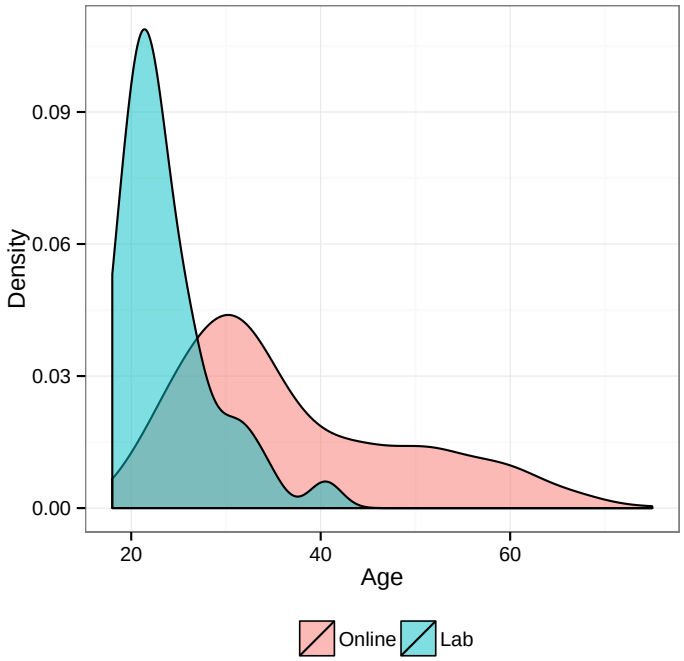


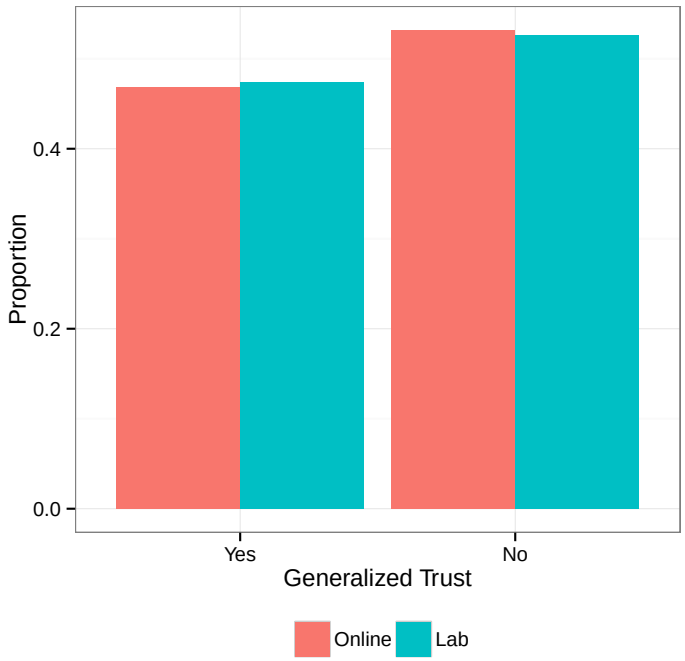
Figure 2: Age of Subjects



Behavioural Measures. As we pointed out earlier, both the online and lab experiments included similar strategies for recovering underlying preferences that might be the source for heterogeneity in redistributive tax preferences. Comparisons of these different metrics across the two subject pools suggest no significant differences. We begin by comparing measures of trust and other-regarding preferences across the subject pools.

First, the level of general trust exhibits the similarity is the lab and online. We use a standard generalized trust question in which we ask a binary question: "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?", and the alternatives are "Most people can be trusted" and "You can never be too careful when dealing with others" (Figure 7). There is essentially no difference between online and lab subjects.

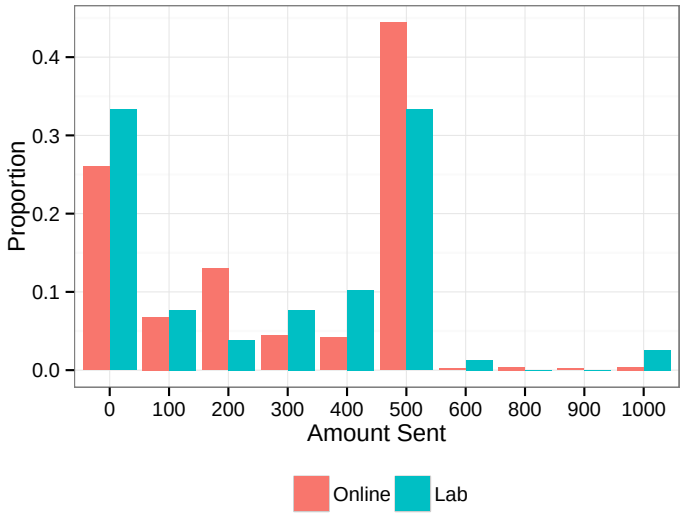
Figure 3: General Trust



We employ the classic Dictator Game described earlier to measure other-regarding preferences. In both the lab and online versions of the Dictator Game subjects have an

opportunity to split an endowment of 1000 ECUs between themselves and a recipient. Figure 6 describes the allocation of ECUs to the recipients. A large proportion of subjects either allocate nothing or a half of the endowment to the recipients. The average amount allocated to the recipient is 299 for the lab, and 283 for the online. And in both t-test and Wilcox rank sum test, the difference between two groups is insignificant. Hence there is little evidence of significant differences in other-regarding preferences between lab and online subjects.

Figure 4: Dictator Game



One could imagine that an important personality trait shaping tax compliance is an individual's underlying integrity. The integrity measure described earlier, which consists of 10 items, was administered to both lab and online subjects. As Figure 5 indicates, lab subjects have significantly higher average scores than online subjects for 6 out of the 10 items – higher scores indicate higher level of dishonesty. A summation of these ten scores results in an overall measure of integrity. Figure 6 shows the distribution of this total score which ranges from 10 (highest integrity) to 40 (lowest integrity). There are a small number of "straight-liners" amongst the online subjects (16 subjects at Score 10 and 1 subject at Score 40). The distribution of scores suggests lower integrity (higher scores) for

the lab subjects. And in fact the average integrity score is higher for lab subjects (19.0) than online subjects (17.3). The difference between the two groups is still significant after applying matching method to recover the balance between groups in term of age and gender.

Figure 5: Integrity Test: Individual Items

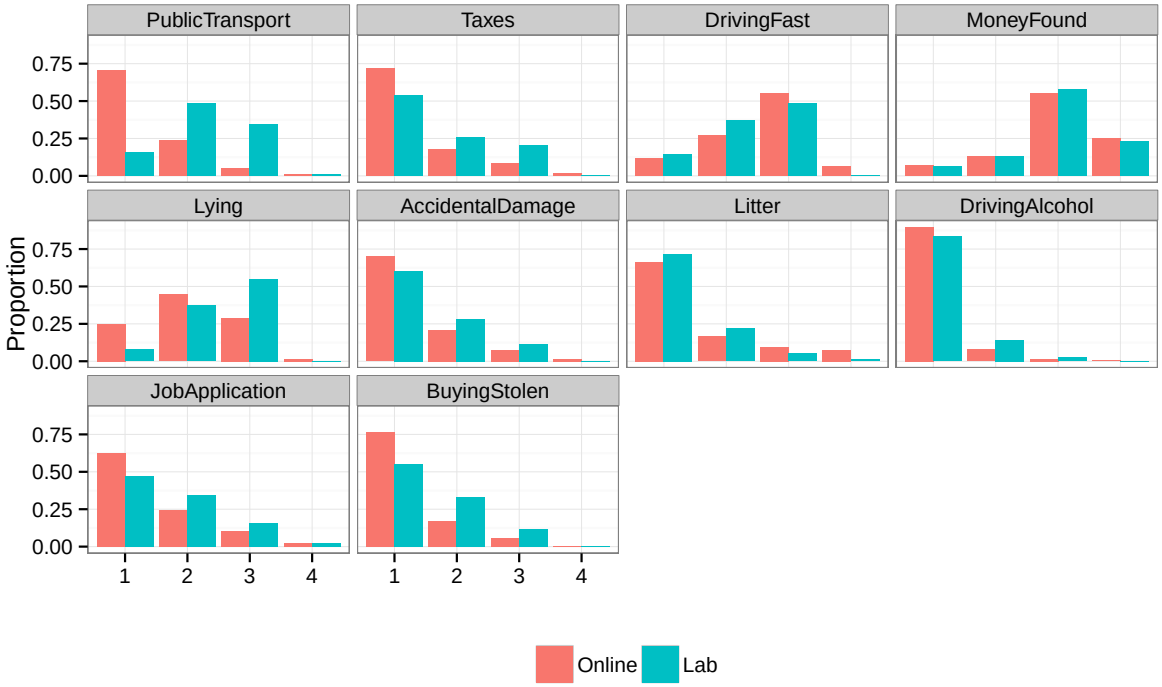
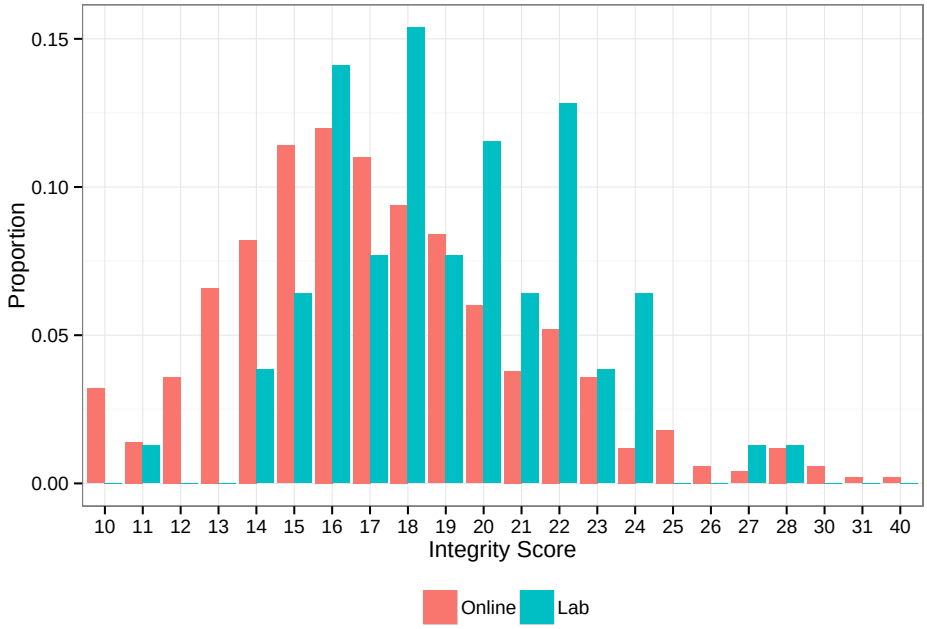


Figure 6: Integrity Test: Total Score

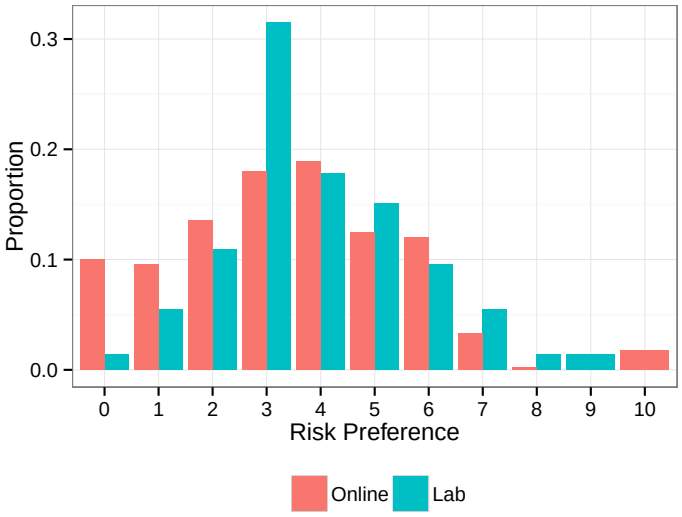


Risk preferences are hypothesized to be strongly correlated with redistributive tax preferences (Duch and Rueda 2013). One argument suggests that those who are risk averse are more likely to favour the redistributive taxation because they see social benefits as a form of insurance. The risk preference elicitation game described earlier is designed to measure these preferences in both lab and online subjects. The measure assumes transitive preference and monotonically non-decreasing utility in terms of monetary earnings. If a subject chooses Option B in a particular lottery (e.g. Lottery 4 in Table 2), then in subsequent lotteries she should choose Option B. Violation of transitivity is often observed in these games. In this experiment, 5 out of 66 lab subjects and 49 out of 500 online subjects exhibit such inconstancy in the test. The rate is slightly higher for online subjects, but the difference is not statically significant. On balance violation of transitivity represents less than 10 percent of subjects – we exclude these subjects from the analysis.

Figure 7 shows the distribution of risk preference after deleting these inconsistent results. The x-axis indicates the risk preference measured as the lottery choice switch. For example the score 4 means that a subject has switched their choice from Option A to B at Lottery 4. The score 0 represents the most risk seeking score while 10 represents the most risk aversive score.[2] Online subjects are slightly more risk seeking but on balance the two subject pools are quite similar with respect to risk preferences.

[2]The score 10 is a logically inconsistent choice because a subject with this score has never switched their choice even if in Lottery 10 the Option B provide a higher earning with a probability of 1. There are eight online subjects with a score of 10.

Figure 7: Risk Preference



3.2 Difference in the demography and behavioural measures

Real Effort Performance. Recall that the tax compliance experiment required subjects to perform a real effort tasks that consisted of adding two randomly generated two-digit numbers in thirty seconds. Figure 8 shows the distribution of outcomes for both lab and online subjects. Online subjects performed slightly lower than did the lab subjects: Average preliminary gain for online subjects is 1715 (11.4 collect answers), while average preliminary gain for lab subjects is 1817 (12.1 collect answers). Lab subjects are younger and, on average, better educated which might explain the slightly higher average. The online version of the tax compliance experiment required us to match, based on the Preliminary Gains, online subjects with lab subjects. Figure 9 is a scatter plot of the Preliminary Gains by online subjects against the replaced lab subjects. This simply confirms that the performance of online subjects is very similar to the performance of replaced subjects ($r^2 = .88$). Because of the slightly higher performance of the lab subjects, a large proportion of online subjects are replacing a lab subjects in the lowest tax bracket (Low bracket: 0.457, Middle bracket: 0.275, Highest bracket: 0.268).

Figure 8: Real Effort Task Performance

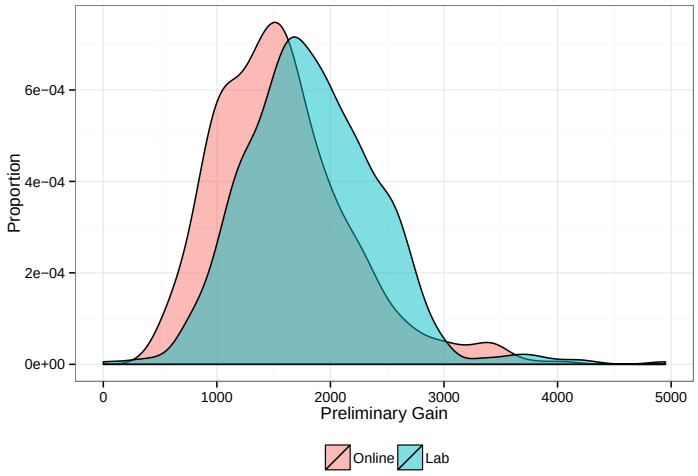
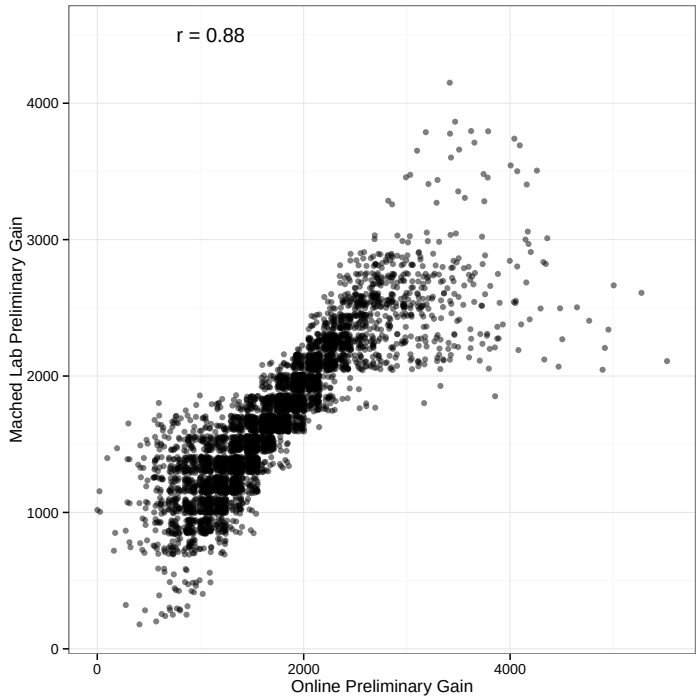


Figure 9: Real Effort Task, Lab-matched Performance



Tax Compliance. One of the innovations of this online experiment is that we leverage existing, similar, lab experiment results in order to simulate real-time interaction amongst the online subjects. We do this by matching the real-effort task outcomes with the previous lab data. In this section, we assess the degree of success of this online experimental setting through an investigation of lab and online data.

In our lab experiment, the subjects are assigned to a six-member group at the start of tax experiment. The lab experiments took place months before the online experiment. In each of the ten rounds of the online experiment, online subjects are matched to one of the lab subjects based on the similarity of their performance on the real effort task. At the income reporting stage the online subjects receives the identical information that was provided to their matched lab subject: the are informed of the tax regime; they learn their tax bracket; and they are informed of the benefits regime which determines how the pooled tax revenues are distributed to each member of the group.

If the online subjects happened to have the identical preference to their matched lab subjects, then the online subjects would make the exact same decision (this controls for treatment effects because the treatment effect is precisely the same in each pairing). In order to test this hypothesis, we matched the replacing subjects for each round with replaced subjects to make a comparable datasets of lab and online results. The matching of lab to online is 1-to-N; in other words, for each data point of lab data, there are "N" multiple points in the online data. Using this matched data, which has 5,000 observations of 500 online subjects who play 10 rounds of tax game.

In this section we mainly compare the income report rate which is the subject's reported earnings divided by the her real effort task earnings. Three online data points in which subjects earned nothing from their real effort tasks are excluded from the analysis. Figure 10 shows the direct comparison of lab and online data by plotting the income report rate of lab against the income report rate of online for replacement pairs. If the

lab and online subjects behaved similarly, the points would be lining up on the 45 degree line. This is clearly not the case. The heterogeneity among subjects makes the direct comparison within replacement pairs implausible.

Figure 10: Plotting Income Report Rates

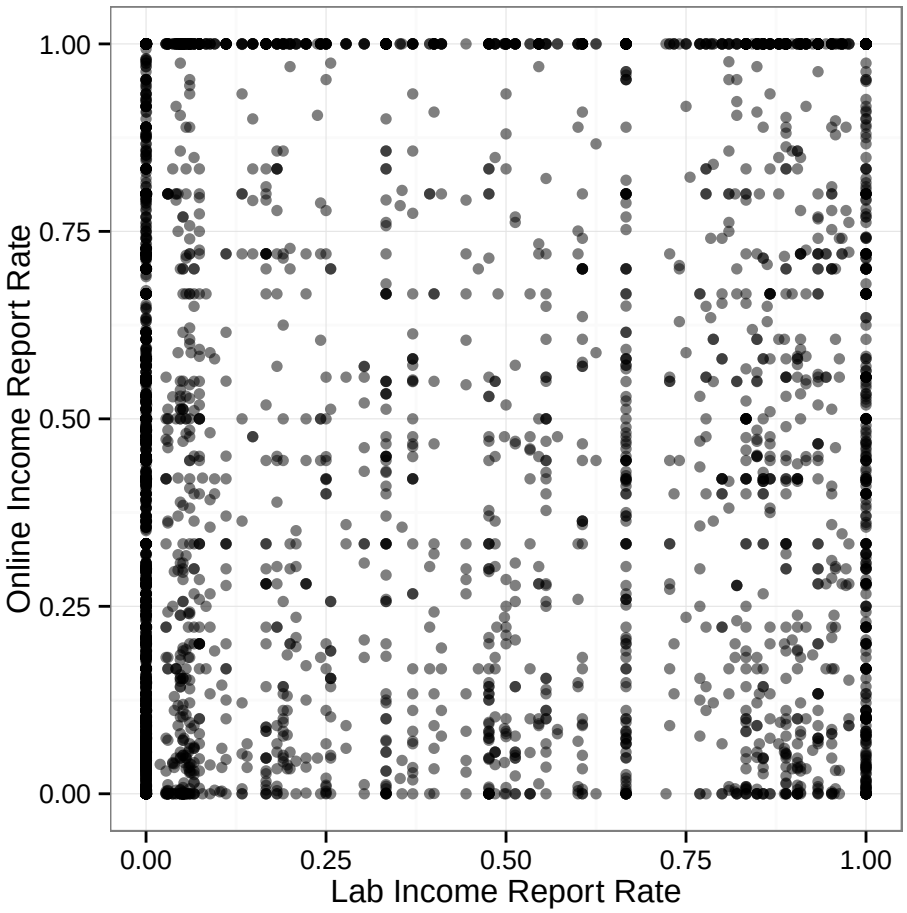
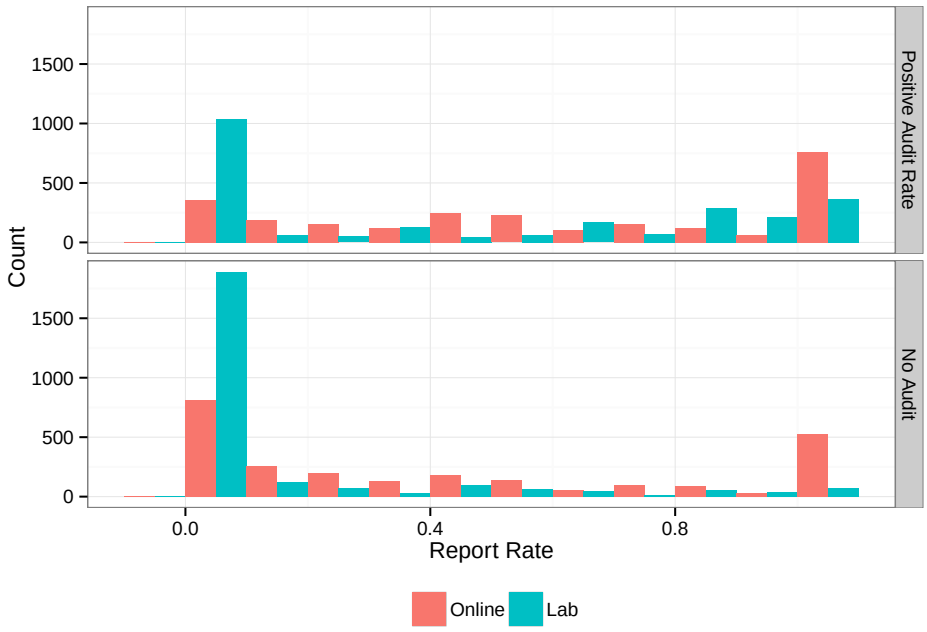


Figure 11 compares the lab and online income report rates using the matched dataset. Reporting no income always maximized subjects' expected earnings in this tax experiment – in both the audit rate treatments. The income report rate takes the value of one in a substantial number of online observations. With the positive audit rate case, about 30.5% of online subjects reported the full income while 14.7% of lab subjects reported full income. These proportions go down when there is no auditing of subjects' income; nevertheless, still 20.9% of online subjects reported their full income while 2.8% of lab subjects did so. For both lab and online subjects, the decrease is about 10%.

Figure 11: Comparison of Income Report Rate



Subjects in both experiments were assigned to similar tax and audit treatments. The results presented in this paper are limited to these fairly broadly defined tax and audit treatments. Beramendi and Duch (2014) report the actual substantive results from the experiments that focus on treatment effects associated with different levels of progressivity related to both revenue collection and the distribution of benefits. The analyses in this paper use the lab and online subjects matched dataset that was described earlier.

Our general expectation is that tax compliance will drop as tax rates rise and that tax compliance will be lower when there is no auditing of income. In the first set of models, we regress dummy variables of tax brackets and auditing rate on the lab or online income report rate. We include two dummy variables for tax brackets, Middle and Low, as well as a "No Audit" dummy variable. The baseline is a High Tax Bracket and a 10% audit rate. For both online and lab models, all estimated coefficients are highly significant in the expected direction. The are some notable difference between lab and online. First, the effect of no audit is larger for the lab subjects. For lab subjects, moving from the no audit treatment to the 10% audit treatment results in an average decrease in reported income of 31 percent. The online subjects are less responsive to the possibility to cheat without retribution. Second, we find similar differences for the tax bracket effects. Lab subjects are more responsive to tax rates. They complying less than online subjects when rates are high and their compliance rates are more responsive to drops in the tax rates.

Behavioral Measures and Tax Compliance

Table 4: Effects of Tax Brackets and Auditing

	Dependent variable:	
	Online Report Rate	Lab Report Rate
	(1)	(2)
Middle Tax Bracket	0.065^{***}	0.094^{***}
	(0.014)	(0.013)
Low Tax Bracket	0.117^{***}	0.138^{***}
	(0.013)	(0.012)
No Audit	-0.173^{***}	-0.311^{***}
	(0.010)	(0.010)
Constant	0.511^{***}	0.354^{***}
	(0.011)	(0.011)
Observations	4,968	4,968
R^2	0.066	0.186
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 5: Effects of Behavioral Preference and Tax Compliance

	Dependent variable:	
	Online Report Rate	Lab Report Rate
	(1)	(2)
Middle Tax Bracket	0.060^{***}	0.043^{***}
	(0.014)	(0.014)
Low Tax Bracket	0.108^{***}	0.096^{***}
	(0.013)	(0.012)
No Audit	-0.183^{***}	-0.301^{***}
	(0.011)	(0.010)
Dictator Game Giving	0.313^{***}	0.227^{***}
	(0.026)	(0.019)
Integrity Score	-0.372^{***}	-0.226^{***}
	(0.043)	(0.052)
Risk Preference	-0.049^{*}	-0.213^{***}
	(0.026)	(0.032)
Constant	0.548^{***}	0.470^{***}
	(0.021)	(0.024)
Observations	4,460	4,542
R^2	0.134	0.217

*p<0.1; **p<0.05; ***p<0.01
Outcome variable: Report rate
Input variables are rescaled to [0,1]

4 Discussion

Lab experiments provided Beramendi and Duch (2014) an ideal vehicle for exploring how the distributive incidence of the public sector over the long run drives the level of tax compliance experiment. Their lab results were compelling but they were concerned with whether their conclusions might be limited given heterogeneous treatments effects and a homogeneous student subject pool. But moving from the experimental lab to online experiments with a much more diverse subject pool raised a number of challenges. In particular, it required a solution to the fact that scheduling real time interactions amongst online subjects is in practice very difficult. The solution was to exploit the existing experimental lab results. They matched online subjects to lab subjects based on the similarity of their performance on real effort tasks. This allowed the experimenters to calculated taxes and redistribute tax revenues to the online subjects.

The goal of this essay is to explore whether there are any mode affects associated with a lab versus online experiment of this nature. First we simply compare the characteristics of the two different subject pools – the student subject pool for the lab experiment was, as we expected, younger. There were no significant differences with respect to gender. Second, we compare the two subject pools in terms of measures relating to underlying personality traits and preferences. Here we do find differences that are in general consistent with the findings of Belot, Duch and Miller (2009) who conduct similar comparisons of student and non-student subject pools. They find that students tend to be less trusting and exhibit lower levels of other-regarding preferences. We find that students subjects, compared to online subjects, show similar levels of trusting and other-regarding preferences, but score lower on a measure of integrity. In short student subjects tend to more closely resemble homo-economicus than is the case with online subjects.

A third issue we explored was whether the two subject pools would perform differently on the real effort tasks that were a critical component of the tax compliance experiments.

The younger and better educated lab subjects performed somewhat better on the real effort tasks. But the differences were not dramatic suggesting that the online subject pool were attentive to the tasks hence were properly incentivised to perform well.

Most importantly, we were interested in whether compliance with the tax rates differed across the two subject pools. And also whether the two subject pools responded similarly to the tax and audit treatments. Consistent with the earlier findings, students tend to be more utility maximizers (or greedy, if you will) than the online subject pool. For any given tax or audit rate they cheat more than the online subjects. And as tax rates rise or audit rates decline their cheating increases a faster rate than that of the online subject pool. Nevertheless, the directional effect of the tax and audit treatments were similar for both subject pools.

References

Belot, Michele, Raymond Duch and Luis Miller. 2009. "Who Should be Called to the Lab?" CESS, Nuffield College.

Beramendi, Pablo and Raymond M. Duch. 2014. "The Distributive Basis of Tax Compliance." Paper prepared for the Symposium in Celebration of Margaret Levi, University of Washington, Seattle, October 10th, 2014.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58(3):739–753.

Boas, Taylor C and F. Daniel Hidalgo. 2013. "Fielding Complex Online Surveys using rApache and Qualtrics." *The Political Methodologist* 20(2):21–27.

Crump, Matthew J C, John V McDonnell and Todd M Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research." *PloS one* 8(3):e57410.

Duch, Raymond M. and David Rueda. 2013. "The Preferences of the Rich and Poor for Redistributional Taxation: UK and Germany." Nuffield College CESS Working Paper.

Grose, Christian R., Neil Malhotra and Robert Parks Van Houweling. 2014. "Explaining Explanations: How Legislators Explain their Policy Positions and How Citizens React." *American Journal of Political Science* 00(0):n/a–n/a.

Healy, Andrew and Gabriel S. Lenz. 2014. "Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy." *American Journal of Political Science* 58(1):31–47.

Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92:1644–1655.

Horton, John J., David G. Rand and Richard J. Zeckhauser. 2011. "The online laboratory: conducting experiments in a real labor market." *Experimental Economics* 14(3):399–425.

Malhotra, Neil and Yotam Margalit. 2014. "Expectation Setting and Retrospective Voting." *The Journal of Politics* 76(04):1000–1016.

Mason, Winter and Siddharth Suri. 2012. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior research methods* 44(1):1–23.

Morton, Rebecca and Kenneth Williams. 2009. *From Nature to the Lab: Experimental Political Science and the Study of Causality*. Cambridge University Press.

Olea, JLM and T Strzalecki. 2014. "AXIOMATIZATION AND MEASUREMENT OF QUASI-HYPERBOLIC DISCOUNTING." *The Quarterly Journal of Economics* pp. 1449–1499.

Paolacci, G., J. Chandler and P.G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5:411–419.

Rand, David G. 2012. "The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments." *Journal of theoretical biology* 299:172–179.

Tsvetkova, Milena and Michael W Macy. 2014. "The social contagion of generosity." *PloS one* 9(2):e87275.