Experimental Modes and Heterogeneity
DRAFT *

Raymond Duch
Centre for Experimental Social Sciences
Nuffield College
University of Oxford
raymond.duch@nuffield.ox.ac.uk

Denise Laroze
Centre for Experimental Social Sciences
Universidad de Santiago de Chile
denise.laroze@cess.cl

Thomas Robinson
Department of Politics and International Relations
University of Oxford
thomas.robinson@politics.ox.ac.uk

Pablo Beramendi
Duke University
pablo.beramendi@duke.edu

July 17, 2018

*Paper prepared for presentation at the Annual Meeting of the European Political Science Association, Milan, Italy, June 22-24, 2017.

Abstract

We assess whether the results of identical interactive experiments differ depending on whether they are conducted in the lab and on-line. The assessments are based on lying experiments in which subjects earn real money, are subject to a deduction (that is redistributed to other subjects), and can lie about their earnings. The impact of experimental mode on estimated effects depends on the nature of the decision-making experiment. There is little evidence that the estimated effect for decision-theoretic experiments varies across mode or subject pool. In the case of incentivized interactive experiments, the treatment effects from traditional lab experiments are much larger than those from on-line experiments. We suggest iterative non-parametric estimation techniques for assessing the magnitude of heterogeneous treatment effects across these modes.

1 Introduction

Experimental social science employs a diverse selection of modes for implementing experiments. These include traditional lab experiments; a wide variety of field and lab-in-the-field experiments; and on-line experiments that employ very diverse subject pools. There are debates across disciplines regarding the appropriate experimental modes (Camerer, 2015; Levitt and List, 2015; Chang and Krosnick, 2009; Coppock, 2018).

This essay makes two contributions to the debate. First, the impact of experimental mode on estimated effects depends on the nature of the decision-making experiment. Consistent with others (Berinsky, Margolis and Sances, 2014; Coppock, 2018; Kam and Simas, 2010; Broockman, Kalla and Sekhon, 2017), we find little evidence that the estimated effect for decision-theoretic experiments varies across mode or subject pool. On the other hand, in the case of incentivized interactive experiments (that we believe are more cognitively demanding for subjects), there are mode effects – in particular, treatment effects from traditional lab experiments are larger.[1]

Second, we illustrate how to leverage diversity in experimental modes to assess the robustness of treatment effects – particularly in the case of incentivized interactive experiments. There is considerable concern across the social sciences with the fragility of estimated treatment effects, and their reproducibility (Maniadis, Tufano and List, 2014; Levitt and List, 2015; Collaboration, 2015). To increase confidence in the robustness of reported treatment effects, we advocate the implementation of identical experimental protocols employing diverse experimental modes. We suggest iterative non-parametric estimation techniques for assessing the magnitude of heterogeneous treatment effects across these modes.

We conduct four identical interactive experiments. One experiment consists of 6 sessions with 116 subjects in the Nuffield Centre for Experimental Social Sciences (CESS) Lab. A second identical experiment is conducted on-line with 144 subjects from the same CESS lab

[1] All of the replication material for this essay is available at: https://github.com/rayduch/Experimental-Modes-and-Heterogeneity

subject pool. In a third experiment 91 subjects from the CESS UK On-line subject pool took decisions in the identical interactive experiment. Finally, 390 MTurk workers, all from the U.S., made choices in the identical interactive experiment.

2 Subject Pools and Experimental Modes

We summarize in Table 1 recent studies that have advanced our understanding of how treatment effects vary by experimental mode.

Decision Theoretic. Virtually all of the studies identified in Table 1 concern experimental treatments in which subjects make decision theoretic choices. Respondents make a static one-shot decision that has no affect on other subjects. And in most cases subjects' decisions are not incentivized. Most of the comparisons of decision-theoretic outcomes reported in Table 1 suggest quite small differences in effect sizes across experimental modes.

Table 1: Evidence on Experimental Mode Effects

Authors	**Subject Pool/Mode**	**Subject Pool**	**Mode**
Kam and Simas (2010)	Lab versus on-line results are similar		
Berinsky, Huber and Lenz (2012)	Lab versus national probability sample results are similar		
Crump, McDonnell and Gureckis (2013)	Lab versus on-line sample results are similar		
Clifford and Jerit (2014)			Lab versus on-line psychological measures results are similar
Mullinix et al. (2015)		Compare framing experiment results for convenience samples versus representative sample and finds little difference	
Belot, Duch and Miller (2015)		Student versus non-student sample results for incentivized interactive experiments are different	
Huff and Tingley (2015)		Compares MTurk and CSES on-line subjects and finds differences in characteristics	
Briones (2017)	Lab versus on-line psychological measures results are similar		
Arechar, Gächter and Molleman (2018)	Lab versus on-line sample results for incentivized interactive experiments are similar		
Coppock (2018)		Compares framing experiment results for large number of different on-line convenience and representative samples and finds little difference	

Most comparisons in Table 1 are of experimental modes that do not distinguish between subject pool and mode effects. Berinsky, Huber and Lenz (2012) present evidence that many of the political preferences and behaviors reported by subjects in the lab and in national probability samples are similar.[2] Kam and Simas (2010) conduct similar framing and risk preferences experiments with students in an experimental lab and with on-line subjects. The treatment effects and recovered risk preferences were very similar in the two modes. Crump, McDonnell and Gureckis (2013) compare treatment effects for a number of primarily cognitive psychology experiments that incorporate reaction times either on the response side or on the treatment side. They find strong similarities in the qualitative effects in the lab and on-line. Again, focusing on psychological questionnaire measures, in this case of sleep quality and stress, Briones (2017) finds that the correlations for students in lab experiments were similar to those for subjects on-line (including MTurk).

Evidence from decision theoretic survey experiments suggests these subject pool differences have little impact on estimated experimental treatment effects. Mullinix et al. (2015) conduct a comparison of treatment effects resulting from framing experiments conducted with a variety of convenience samples (including MTurk and student subjects) with those from identical experiments conducted on-line with nationally representative samples of the U.S. population. The estimated treatment effects are surprisingly similar across the different on-line experiments with quite different subject pools.

As Table 1 suggests, the composition of subject pools seems to have little bearing on the outcomes of decision theoretic experiments. And this is in spite of the fact that most on-line subject pools fall short of the representativeness of national probability samples (Chang and Krosnick, 2009; Huff and Tingley, 2015). Coppock (2018) provides strong evidence of a robust correlation between the treatment effects of two types of on-line subject pools: Mturk subjects versus representative samples (for the most part of the U.S. national population). This is based on the Mturk replication of 12 studies that were conducted on-line with repre-

[2]One of the comparisons in Berinsky, Huber and Lenz (2012) is MTurk versus the ANESP which is conducted on-line.

sentative U.S. populations. Mullinix et al. (2015) compare framing experiments conducted with a variety of convenience samples (including MTurk and student subjects) with those from identical experiments conducted on-line with nationally representative samples of the U.S. population. The estimated treatment effects are surprisingly similar across the different on-line experiments with quite different subject pools.

There are well-known concerns as to whether subjects in on-line experiments are actually being treated because of low-levels of attention and comprehension (Berinsky, Margolis and Sances, 2014); poor incentives (low wage rates); opportunities to be treated multiple times in the same experiment; accessing information on the Internet; and unauthorized communication with other subjects in the experiment. But evidence to this effect is limited. Clifford and Jerit (2014) administer several psychological measures in parallel lab and on-line surveys and find little difference for these standard psychological outcome measures.[3]

Hence, the bulk of these decision theoretic studies comparing different experimental modes suggest that 1) responses to the questions typically asked in social science surveys – partisanship, political interest, voting turnout – vary across modes but not in a particularly dramatic fashion; and 2) treatments effects for the typical decision-theoretic experiments that we see in social science research are quite similar across modes. The subject pools clearly differ in terms of many key demographics and, while these can be a source of heterogeneous treatment effects, its often the case that estimated effects are quite similar across diverse subject pools. It is reasonable to conclude that with respect to decision theoretic non-incentivized experiments mode or subject pool effects are typically quite similar.

Interaction and Incentives. There is a rich tradition of incentivized experiments in which subjects make decisions in real time that have financial consequences both for themselves and others. The protocols for these experiments are typically complex. In particular, subjects are provided with more information about the consequences of their choices and

[3]The authors do find a significant difference in the in political knowledge measure suggesting that subjects consult with outside sources in on-line setting. And self-reported levels of distraction are much higher for on-line participants.

other real-time factors that shape their earnings in the experiment. And these experiments can be more cognitively demanding in the sense that subjects are digesting streams of updated information in real time. It is more challenging to design and implement these experiments for the on-line mode. And in fact there are relatively limited numbers of these experiments in which on-line subjects play real-time interactive games (e.g. Mason and Suri, 2012; Arechar, Gächter and Molleman, 2018) – most on-line experiments do not use real-time interaction.[4]

Here mode likely matters. For these more complex and demanding incentivized interactive experiments, our expectation is that traditional lab environments have an advantage. These advantages, related to internal validity, are well known (Morton and Williams, 2009). In fact many of the typical drawbacks of decision-theoretic experiments implemented on-line are exaggerated for interactive online experiments: its difficult to randomly assign subjects to groups in real time; we do not observe who is being treated; subjects are distracted; and attrition is a serious problem. Hence at the very least we might expect noisier and possibly biased estimated treatment effects.

We have relatively little evidence regarding mode and subject pool effects for these incentivized interactive experiments. The exception is Arechar, Gächter and Molleman (2018) who compare identical real-time incentivized interactive experiments conducted on-line with MTurk subjects and in the lab with student subjects. They find little difference in treatment effects. And while their design makes it difficult to disentangle mode and subject-pool effects, the null result certainly suggests that compliance with treatment may, contrary to our expectation, be quite similar on-line and in the lab.

Leveraging Heterogeneity. Heterogeneous treatment effects are problematic if they are simply an artifact of the experimental mode or subject pool we employ. This is strong evidence for fragile treatment effects. Determining whether mode or subject pool is a source

[4]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).

of heterogeneity contributes to establishing the robustness of an hypothesized treatment effect (Neumayer and Plumper, 2017).

There is a growing literature on iterative machine learning methods for estimating heterogeneous treatment effects (Athey and Imbens, 2017; Grimmer, Messing and Westwood, 2017). They allow us to estimate the magnitude of treatment effects for all possible combinations of relevant covariates. This estimation strategy combined with an experimental design that employs diverse experimental modes is a useful method for assessing the robustness of treatment effects. We illustrate with an experimental design that has identical experiments implemented in four diverse experimental modes. And we implement one of a number of iterative machine learning methods to estimate heterogeneity in treatment effects associated with these modes and subject pools.

Illustration. The experiments we describe and analyze here illustrate three broad conclusions from the literature on experimental modes and heterogeneous treatment effects. It is generally recognized that convenience samples will both differ from each other (MTurk will look different than other on-line convenience samples) and likely deviate from a national representative sample. We can speak to the former and demonstrate that demographic covariates differ considerably. But for a broad range of experiments, including ours this convenience sample diversity is not the source of significant treatment heterogeneity.

We distinguish between choices made in decision-theoretic versus incentivized interactive games. Our experiments incorporate both decision-theoretic and interactive games. The literature suggests that decision-theoretic treatment effects will not differ significantly across either subject pools or experimental modes. We illustrate that this is in fact the case by comparing the distributions of decision theoretic responses across the four experimental modes.

On the other hand, in incentivized interactive experiments we expect treatment effects to vary by mode but not by subject pool. The outcome of interest in our principal experiment

is lying. The primary treatment effect of interest is whether subjects are of high versus low ability – and our expectation is that the correlation between ability and lying will vary across modes. We identify heterogeneous treatment effects with automated statistical learning methods. We combine the four experimental results and use FindIt and BART to estimate the full set of conditional average treatment effects (CATE). Our expectation is that the CATEs would suggest that much of the heterogeneity in treatment effects is associated with associated with experimental mode.

3 Experiments

The Design. We implement similar protocols to the Duch, Laroze and Zakharov (2018) lying game in which subjects earn money performing real effort tasks (RET); deductions are then applied to their earnings and distributed to other group members (subjects are randomly assigned to groups of four); and subjects have opportunities to lie about their earnings. In all experiments, subjects make the same interactive decisions in real time. We implement four identical experimental protocols employing four different types of experimental conditions. These are summarized in Table 2. Our goal is to leverage this design to gain insights into subject pool and experimental mode effects.

Table 2: Summary of Mode and Subject Pool Effects

		Mode	
Subject Pool	Lab	On-line	Mode Effects
Lab Student	Yes	Yes	Estimated
On-line (M-Turk)	No	Yes	NA
On-line (CESS UK)	No	Yes	NA
Subject Pool Effect	NA	Estimated	

The rows of Table 2 indicate how subject pools differ in our design and the columns indicate how modes differ. The first row of Table 2 indicates that we observed student

subjects from the CESS Lab subject pool making decisions in the lying experiment. Moreover the two mode columns indicate that we observe these subjects playing the game both in the lab and on-line. The third column suggests that we have a reasonable estimate of mode effects since we are effectively controlling for subject pool characteristics.

The second row of Table 2 refers to the MTurk version of the lying game. Note that we only observe the MTurk subject pool making decisions on-line so we are not able identify a distinctive lab versus on-line mode effect for MTurk subjects. Similarly for the third row – the CESS UK on-line subject pool. We only observe these subjects on-line – hence no distinctive mode effect estimated for these subjects.

The first column indicates that we only observe subjects from the CESS Lab subject pool participating in the lab experiment. Hence, here we cannot directly estimate a subject pool effect controlling for lab mode. The second column of Table 2 indicates that we observe all three subject pools making decisions on-line. We leverage this to estimate subject pool effects controlling for on-line mode.

Laboratory Experiments. The lab experimental sessions were conducted at Nuffield CESS in Nov-Dec 2013 and Aug-Sep 2017. The experiment consists of five modules, two lying modules (one with and one without auditing), a Dictator Game, a Holt-Lowry risk preferences game, and a non-incentivized questionnaire. In the first three modules, we offer earnings in Experimental Currency Units (ECU). The conversion rate is 300 ECUs to 1 British Pound. Instructions are read out loud before each module. The lab experiment takes on average one and a half hours.

The experiment begins with a Dictator Game. This is followed by two lying modules consisting of ten rounds each and they only differ in the audit rates – 0% audit in the first module and 20% audit in the second. Prior to the lying game, participants are randomly assigned to groups of four and the composition of each group remains unchanged throughout the both lying modules. Each round of these two lying modules has two stages. In the first

stage subjects perform RET to compute a series of two-number additions in one minute. Their Preliminary Gains depend on the number of correct answers, getting 150 ECUs for each correct answer.

In the second stage, subjects receive information concerning their Preliminary Gains and they are asked to declare these gains. A certain percentage of these Declared Gains is then deducted from their Preliminary Gains. These deductions are then summed up and evenly divided amongst the members of the group. Note that in each session the deduction rate is consistent. The deduction treatments implemented in the lab experiments are: 10%, 20% and 30%. Subjects are informed of the audit rate at the beginning of each module and that, if there is an audited discrepancy between the Declared and Preliminary gains, they will be deducted half of the difference between the two values plus the full deduction of the Preliminary gains.

At the end of each round participants are informed of their Preliminary and Declared gains; the amount they receive from the group deductions; and their earnings in the round. Subjects are paid for one out of the ten rounds in each lying module at the end of the experiment, and do not receive feedback about earnings until the end of the experiment.

The lying modules are followed by a Risk Preference Game. And the final module is a questionnaire that measures preferences and socio-demographic characteristics. The demographic variables are gender and income. Preference modules include a measure of trust, integrity, other-regarding preferences and risk aversion. The details of these games are provided in Duch, Laroze and Zakharov (2018) and in their On-line Appendices.

On-line Experiment. We also conduct an on-line version of the lying experiment with three different subject pools – the same student subject pool eligible for the lab, a general population UK panel (CESS on-line), and U.S. Mturk workers. The only substantive differences are that: 1) participants play one cheating module of 10 rounds instead of the two modules that exist in the lab version. The second cheating module is omitted to reduce the

length of the experiment. In the lying module there is either a 0% or 10% audit rate that is fixed throughout the session. 2) There are only on screen instructions. 3) The conversion rate is lower, at 1000 ECUs = £1 for UK samples (US $ for Mturk) (compared to the 1000 ECUs = £1 in the lab).

In line with the lab treatment, the deduction rates implemented in the on-line experiments are 10% and 30%; they are randomly assigned to the entire group; and are constant throughout the ten rounds. Groups are composed of four people and are constant for the entire session. Subjects' are informed of the deduction and audit rates, and potential penalties for cheating, which are the same as in the lab. Subjects are paid for their decisions in the dictator game, the results of one out of ten rounds of the cheating module and a risk lottery. There is also a questionnaire measuring trust and socio-demographics.[5]

4 Results: Lab and On-line Experiments

4.1 Sample Covariates

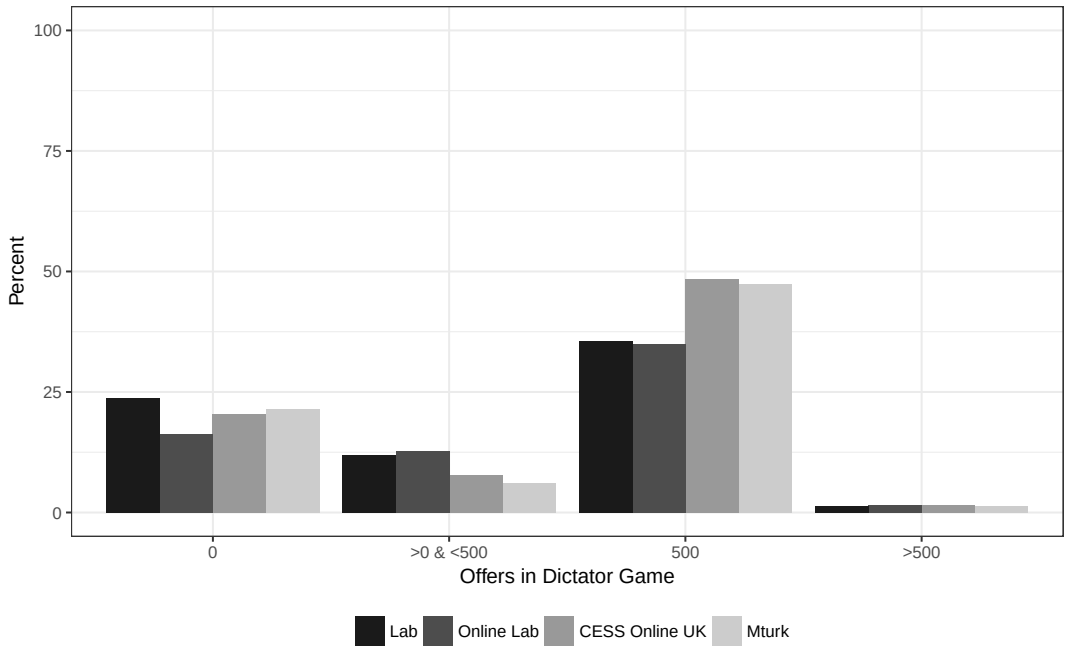
Socio-demographics vary across subject pools. The gender distribution of subjects in the lab and on-line are quite similar except for the UK lab sample, that has a higher proportion of male subjects. There are substantive age differences in the three subject pools. We know that MTurk workers tend to be younger than population survey samples (Berinsky, Huber and Lenz, 2012), and as we would expect, the undergraduate student subjects both in the lab and on-line are even younger on average. The general UK on-line panel subjects are similar to MTurk subjects. The age distributions for MTurk and UK on-line are significantly different from UK lab and on-line in both t-test and Wilcoxon rank sum test, but MTurk and UK on-line are not distinguishable at the 95% confidence level.

[5]In the on-line version we also incorporated a die game as a second measure of cheating. Those results are not analyzed for this study as there are no comparative data from the lab.

4.2 Decision-theoretic preferences

Our subject pools may differ with respect to fundamental preferences – a concern, for example, raised in Belot, Duch and Miller (2015). We implemented a set of incentivized decision theoretic experiments designed to recover a number of standard preferences.

Figure 1: Dictator Game



Other-regarding preferences are similar across the different subject pools but there are differences. We employ the classic Dictator Game to measure other-regarding preferences. In both the lab and on-line versions of the Dictator Game subjects have an opportunity to split an endowment of 1000 ECUs between themselves and an undisclosed recipient. Figure 1 describes the allocation of ECUs to the recipients dividing the subjects into those that gave nothing to the other person, gave something but less than half, those that split the ECUs evenly and those that gave more than half. 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 286 by students in the lab, 303 by students on-line, 329 by the general UK panel and 307 by Mturk workers.

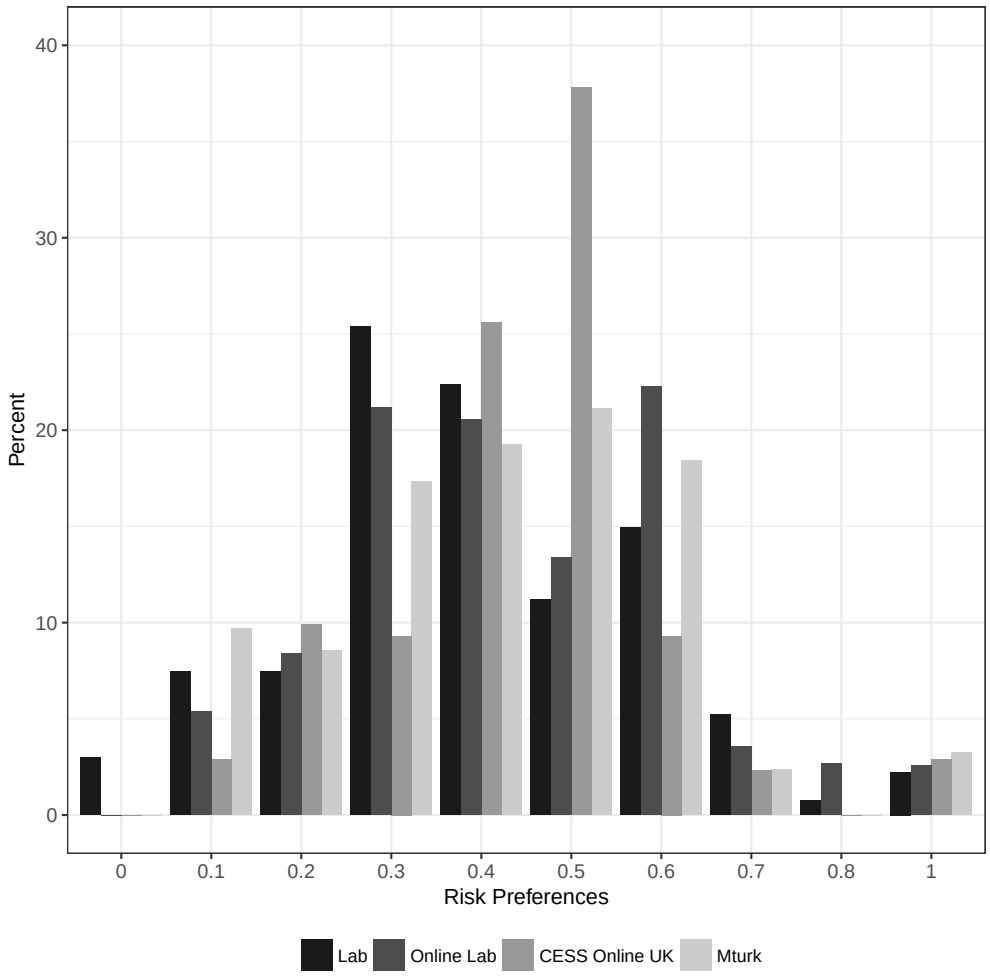
Students are more likely to offer nothing when they are in the lab, but in both t-test and Wilcox rank sum test, the difference between students in the lab and on-line is insignificant. In contrast, the UK On-line panel and Mturk subjects are significantly more generous than the two student subject pools. This is confirmed by both t-test and Wilcox rank sum tests. Mturk workers and participants in the CESS on-line UK panel are indistinguishable from each other.

A second incentivized experiment elicited the risk preference of both lab and on-line subjects employing a standard Holt-Lowry instrument.[6] Participants were asked to make ten choices between two lottery's Option A (less risky) and Option B (more risky). In expectation pay-offs are higher for Option A for the first four decisions and then Option B has a higher expected pay-off. 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, then in subsequent lotteries she should choose Option B. Violation of transitivity is often observed. In this experiment, most subjects reveal consistent preference, with inconsistency ranging from 13 percent of lab students on-line, 16 percent of students in the lab, 17 percent of Mturk workers, and, a surprisingly high, 31 percent of CESS On-line subjects. In the following analyses, eliminating these observations does not substantively alter the results. Therefore observations are kept to avoid reducing the sample size.

Figure 2 shows the distribution of risk preference from the studies. The x-axis in Figure 2 presents a ratio of the number of times a participant chose Option B over the total ten decisions. CESS On-line subjects are slightly more likely to score 0.4-0.5, in the risk neutral range, but overall the different subject pools are quite similar with respect to risk preferences. Note that we omitted from the analysis the risk preference observations for people who participated in the on-line versions of the experiment and had a risk preference of zero. These subjects never selected Option B, even when it was certain that Option B paid £1.85 more than Option A. In the on-line experiments, a risk preference of zero could result from

[6]Details on the lottery experiment are provided in the On-line Appendix.

Figure 2: Risk Preference



1) the participant logging off (in those cases the code recorded the answers as zero/Option A); 2) not understanding/reading the instructions. This did not occur in the lab.

Subjects in the lab made less generous offers in the Dictator Game than other subjects. There is weak evidence that this is a mode effect. The lab pool subjects playing the Dictator Game in the lab were significantly less generous than subjects playing the same game on-line (Cess Online UK: $p < 0.001$; MTurk: $p < 0.05$) although the difference between subjects from the same lab subject pool playing the game on-line and in the lab does not reach conventional levels of significance ($p > 0.1$). And at least two of the three different on-line subject pools made very similar average offers in the Dictator Game. On the second

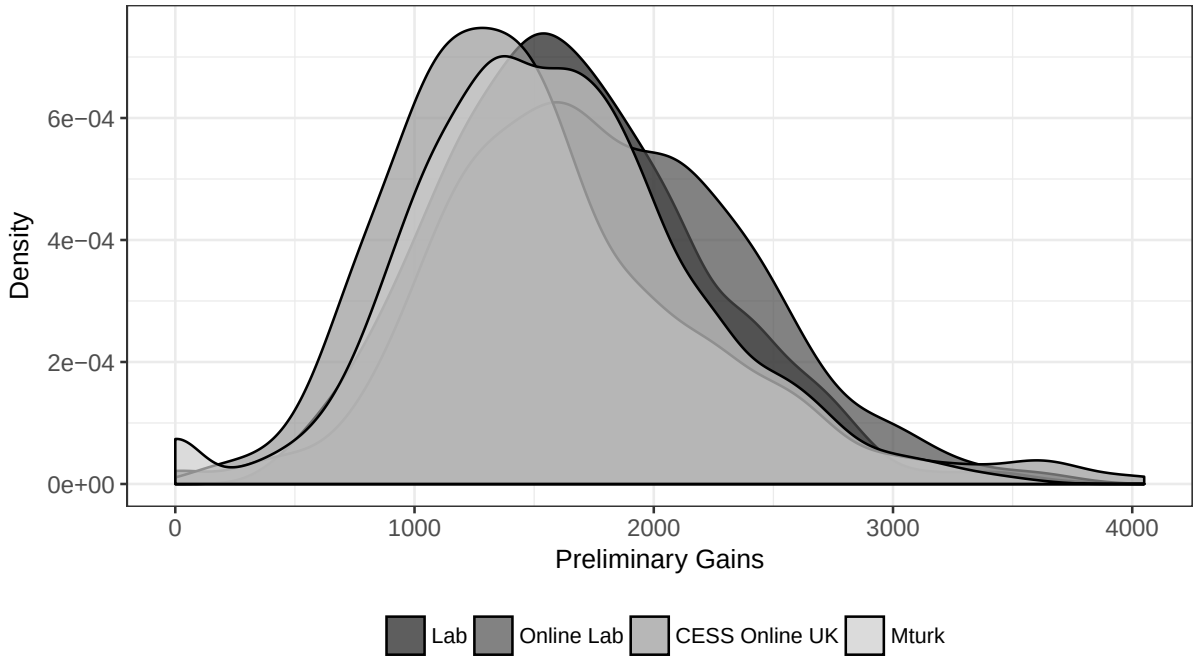
incentivized risk preference experiment subjects made similar choices – none of the risks results from the four experiments suggested a significant mode or sample difference.

4.3 Interactive decision-making

The lying game differs from the decision-theoretic experiments in that subjects had to invest effort to earn money, make decisions about lying, and participated in groups, in real time, that shared income generated from deductions from individual earnings. We view this is a strong test of treatment effect equivalency across subject pools and modes.

Real Effort Performance. In all four experimental modes, subjects were paid to add two randomly generated two-digit numbers in one minute (payment to on-line subjects were lower than in the lab). Figure 3 shows the distribution of outcomes for both lab and on-line subjects. Despite minor variations in the distributions, there are no substantive differences in average gains across subject pools or modes. The average Preliminary Gains for CESS On-line was 1519 ECU (10.13 correct answers), equivalent to the 1574 ECU (10.50 correct answers) obtained by Mturk workers. Students, on average, obtained 1659 ECU (11.06 correct answers) in the lab and 1775 ECU (11.85 correct answers) on-line. Student subjects (Lab and On-line) are primarily Oxford undergraduates and are, on average, better educated which might explain the higher performance. MTurk subjects performance is slightly higher than UK on-line, possibly a result of being "professional" on-line workers.

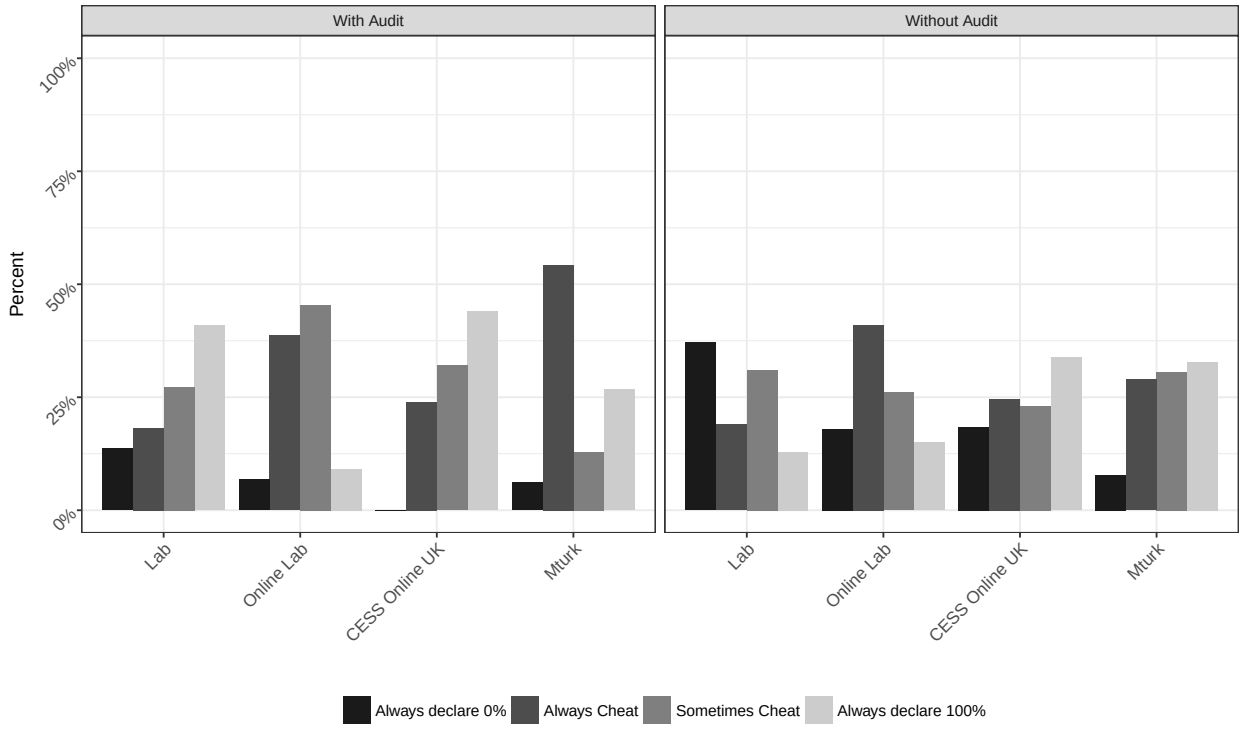
Figure 3: Real Effort Task Performance



Lying. Subjects have an opportunity to lie in this experiment. One might imagine, for example, that subjects in a lab experiment will be more hesitant about lying given the presence of the experimenter and other participants in the lab. The anonymity of on-line subjects might exaggerate lying.

The on-line replications of the lying experiment closely resembled the lab versions in that on-line subjects were randomly assigned to groups of four and played the lying game in real time with the other group members. There are two behavioral difference that stand out the zero-audit condition in Figure 4: First, subjects in the lab are much more comfortable lying about their income. Second, and less dramatic, MTurk workers seem more hesitant about lying than is the case for other on-line participants. In the zero audit condition, over 30 percent of MTurk subjects report 100 percent of their income while only 13-15 percent of the CESS lab subjects, in either mode, behave similarly.

Figure 4: Comparison of Income Report Rate



Treatment effects The treatment effects of interest in these experiments are developed and explored in detail elsewhere (Duch, Laroze and Zakharov, 2018). Subjects in all experiments were assigned to similar deduction and audit treatments. Our general expectation is that those who perform better on the RET will lie more about their income; lying will drop as deduction rates rise; and lying will be lower when there is no auditing of income.

Table 3 reports results for the regression model with the percent of income reported as the dependent variable. The independent variable is ability measured by the rank of one's average performance across all experimental rounds relative to all other participants (normalized between 0 and 1, where 1 is the highest performer). We include two dummy variables for the 20% and 30% deduction rates, and a "No Audit" dummy variable. In addition, we include age and gender covariates as further controls. The baseline is the 10% deduction rate and a 10 percent audit rate. For the four on-line and lab models, the estimated coefficients for Ability Rank are, as expected, negative and significant in all four equations.

The Deduction dummy coefficients are negative and significant for the Lab subject pools (but not for the on-line subject pools). And the Audit dummy variable is negative and significant in three of the four models. The lab results stand out as being most consistently supportive of our conjectures. The experiments outside of the lab are less consistently supportive although again not contradictory.

Table 3: GLM estimation on percent declared

	Mode			
	Lab	Online Lab	Online UK	Mturk
Ability Rank	-0.500^{***}	-0.163^{***}	-0.163^{**}	-0.120^{***}
	(0.036)	(0.045)	(0.071)	(0.037)
20% Deduction	-0.123^{***}			
	(0.024)			
30% Deduction	-0.128^{***}	-0.184^{***}	0.042	0.018
	(0.025)	(0.025)	(0.038)	(0.021)
No Audit	-0.334^{***}	-0.127^{***}	-0.155^{***}	0.011
	(0.023)	(0.026)	(0.036)	(0.024)
Age	0.012^{***}	0.007^{**}	-0.0002	0.002^{**}
	(0.002)	(0.003)	(0.001)	(0.001)
Gender	0.002	0.100^{***}	-0.022	-0.004
	(0.022)	(0.025)	(0.035)	(0.020)
Constant	0.715^{***}	0.476^{***}	0.880^{***}	0.576^{***}
	(0.066)	(0.089)	(0.070)	(0.043)

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered by participant

The estimated standard errors clustered by participant in Table 3 understate the uncertainty for the on-line experiments given the relatively small number of subjects in each of

Table 4: Wild and PCB clustered p-values

	Wild				**PCB**			
	Lab	Online Lab	Online UK	MTurk	Lab	Online Lab	Online UK	MTurk
Constant	0.00	0.07	0.00	0.00	0.00	0.07	0.00	0.00
Ability Rank	0.00	0.21	0.46	0.22	0.00	0.20	0.45	0.22
20% Deduction	0.11				0.10			
30% Deduction	0.12	0.01	0.73	0.76	0.13	0.02	0.71	0.77
No Audit	0.00	0.07	0.11	0.84	0.00	0.06	0.15	0.81
Age	0.06	0.48	0.95	0.49	0.07	0.48	0.95	0.46
Gender (1 = Female)	0.98	0.19	0.84	0.95	0.98	0.16	0.84	0.95

these experiments (Esarey and Menger, 2018). Accordingly, Table 4 reports p-values for the coefficients in the GLM models using wild cluster bootstrapped t-statistics ("Wild") and pairs-clustered bootstrapped t-statistics (PCB) respectively (Cameron, Gelbach and Miller, 2008). The significance of our coefficient estimates diminish substantially as we would expect. Despite this, the significance of one's ability remains highly statistically significant in the lab across both clustering procedures. The No Audit condition is significant in the lab setting, and marginally significant for on-line lab participants. 30 % deduction rates also remain significant ($p < 0.05$) for on-line lab participants.

The estimated effects reported in Table 3 are significant and in the expected direction for the lab experiments; there is more variability in direction and significance for the on-line multivariate results. The Wild and PCB p-values for the coefficients reported in Table 4 indicate that effects for on-line experiments are much more imprecisely estimated. There is a suggestion here of heterogeneous treatment effects related to experimental mode.

Heterogeneity. The coefficients on ability are consistently negative, but imprecisely estimated in some cases; the magnitudes of the ability coefficients vary; and demographic covariates are significant in some, although not all, models. Should we conclude the estimated ability effect is robust across modes and subject pools? We propose a procedure

that identifies heterogeneous effects that are associated with modes or subjects. It estimates conditional average treatment effects (CATE) for the combined data from the four identical experiments. The estimates provide two insights into robustness. First, is there evidence in the CATEs contradicting the conclusions regarding average treatment effects estimated in Table 3? Second, are there significant differences in CATEs across experimental modes?

The challenge is to identify which interactions of covariates and treatment effects are noteworthy here. As many have pointed out, there are significant advantages to automating this estimation by employing non-parametric iterative estimation techniques (Green and Kern, 2012; Imai and Strauss, 2011; Athey and Imbens, 2017; Grimmer, Messing and Westwood, 2017). These techniques allows us to assess whether experimental modes condition estimated treatment effects (Green and Kern, 2012; Imai and Strauss, 2011). Our strategy is to estimate CATEs for subjects who share particular values on all combinations of relevant covariates including the experimental modes in which they participated.

In spite of the relatively large numbers of subjects in these experiments, the number of subjects populating any one unique covariate/mode values will be relatively small. With so few observations sharing any one of these unique covariate values, estimated differences in CATEs are likely to be driven by random variation in the small samples (Athey and Imbens, 2017; Grimmer, Messing and Westwood, 2017). The challenge then is to estimate heterogeneous effects that distinguish systematic responses from differences that are the result of chance random assignment.

A number of techniques have been proposed for overcoming this limitation in estimating the response surface for any treatment variable conditional on particular covariates. Grimmer, Messing and Westwood (2017) present an excellent overview along with suggestions for estimating a weighted ensemble of such estimators. We present the results of two automated statistical learning estimation strategies.

Imai and Ratkovic (2013) suggest one strategy employing a Support Vector Classifier (SVC). The iterated LASSO model estimates produced by the Imai and Ratkovic (2013)

algorithm result in an average treatment effect for each combination of values for the specified vector of covariates hypothesized to be the source of heterogeneity. Of interest here is whether our two experimental conditions – on-line versus lab mode and student versus non-student subject pools – are a significant source of heterogeneity in the treatment effects.

We first estimate a complete interactive model specification with student and on-line dummy variables, as well as age and gender covariates (also included in the interaction).[7] In line with Imai and Ratkovic (2013), this model is initially fitted through a series of iterated LASSO fits that result in optimal estimates of the LASSO tuning parameters. The model incorporates separate LASSO constraints for the treatment effect heterogeneity variables (λ_Z) and the remaining covariates in the model (λ_V). A final estimate of the model coefficients for the ATE (Ability Rank) and interactive effects is generated using the converged values of the LASSO tuning parameters.[8]

In our case, the LASSO model generated non-zero heterogeneous parameter estimates for subjects within both mode conditions. This is particularly noteworthy given the sparse estimation strategy of LASSO models. Conditional on each subset of covariate values, designated as the source of heterogeneous treatment effects, a CATE is generated by taking the difference of predicted outcomes in both treatment and control for the subset of subjects. From this model, we predict the expected effect of treatment for each individual's sample, mode and treatment assignment plus their vector of covariate values.

[7]We estimate this model using the FindIt package within R. See Egami, Ratkovic and Imai (2018) for further details on the package specification and procedure.

[8]Each iteration of the LASSO fit is conducted on a subset of the full sample, and is thus a cross-validation procedure. Optimization of the LASSO constraints is achieved through an alternating line search that attempts to minimise a generalized cross-validation statistic. Imai and Ratkovic (2013) provide a detailed discussion and full specification for the GCV statistic used.

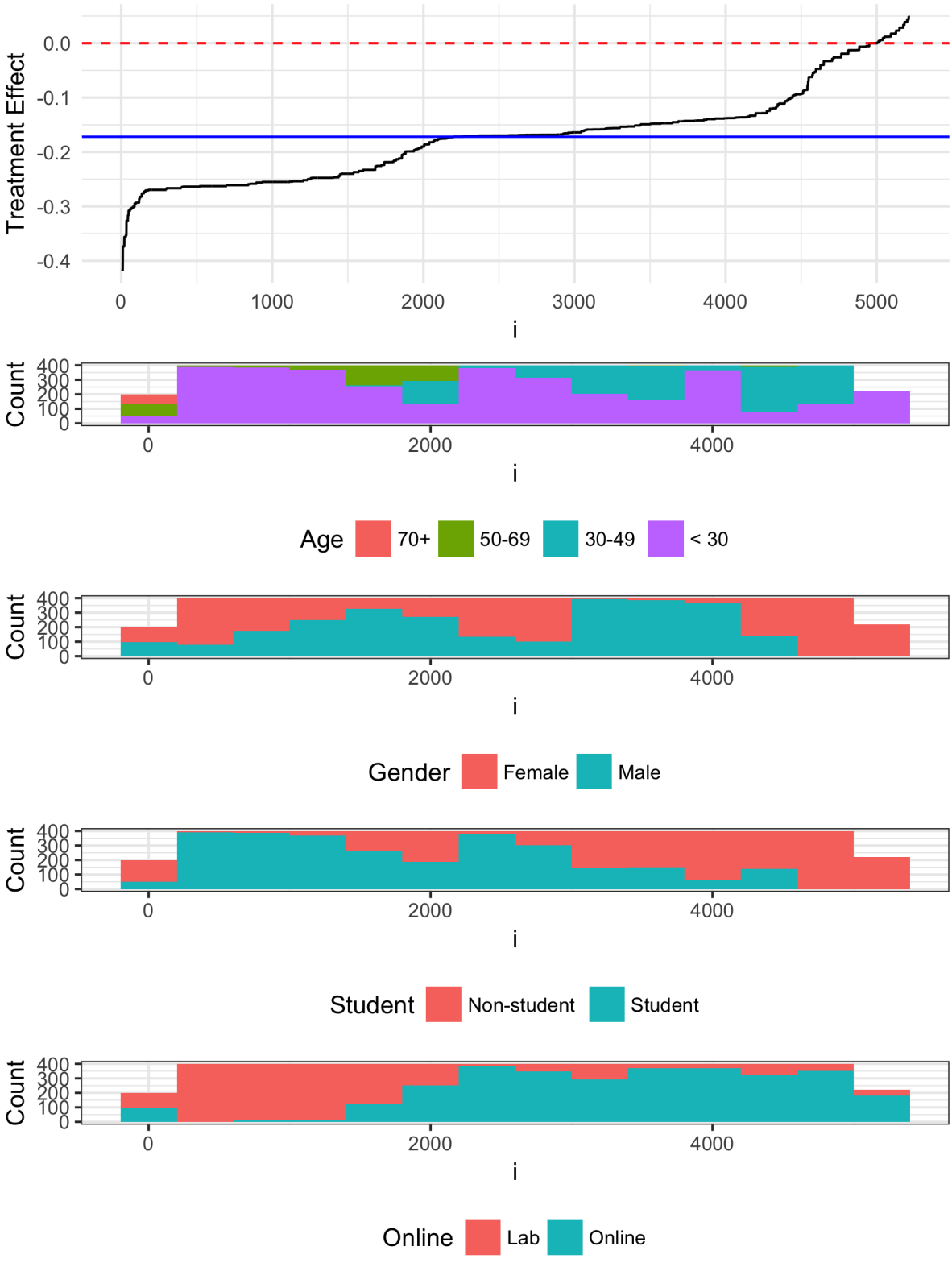
Table 5: Heterogeneous treatment coefficients and interactions using iterated LASSO model

Variable	Coefficient
Student	-0.066
On-line	-0.054
Age	0.006
Gender	0.029
Student × On-line	0.059
Student × Age	0.002
Student × Gender	0.206
On-line × Age	0.001
On-line × Gender	0.032
Age × Gender	0.003
Treatment	-0.133
Treatment × Student	-0.087
Treatment × On-line	0.047
Treatment × Gender	0.033
Treatment × Student × On-line	0.059
Treatment × Student × Age	0.006
Treatment × Student × Gender	-0.186
Treatment × On-line × Age	-0.002
Treatment × On-line × Gender	0.005
Treatment × Age × Gender	-0.005
Intercept	0.602
ATE	-0.172

A CATE is estimated for each subject based on the model presented in Table 5 and their individual vector of treatment and covariate values. Figure 5 summarizes the estimation results. The horizontal blue line indicates an overall ATE of -0.172. The individual estimated heterogeneity effects are organized such that the largest negative effect is on the left while the extreme right represents estimated CATEs that approach zero – there are a few that in fact exceed zero. Recall that the expected effect is negative.

The lower part of Figure 5 presents the counts of the subject choices corresponding to the range of treatment effects in the upper part of Figure 5. Each of the four histograms provides either a covariate or mode/sample pool profile of subjects at each of the treatment effect values. Almost all of the subjects who played the game in the lab are concentrated at the extreme negative tail of the CATE distribution. And subjects who played the game on-line are for the most part distributed to the right of those taking decisions in the lab, i.e. on-line CATEs have lower magnitudes. Most interestingly, this includes student lab subjects who played the game on-line. As the student/non-student pool histogram indicates, they also had CATEs stretching well to the right on the distribution. Hence they more closely resembled other subjects playing the game on-line rather than their fellow-students playing the game in the lab.

Figure 5: FindIt estimated heterogeneous effects including covariate interactions



It is clear from the histograms that students in the lab account for an important heterogeneous effect – the Ability Rank effect is particularly strong and negative for these subjects. What is striking, and something that was hinted at in the multivariate analysis, is that mode is a source of heterogeneity. There are essentially no lab subjects with estimated treatment effects falling above the overall estimated ATE.

A second frequently employed non-parametric modeling strategy is Bayesian Additive Regression Trees (BART) (Green and Kern, 2012; Hill, 2011). This method is a Bayesian adaptation of the frequentist CART strategy for estimating tree models that repeatedly divide up the sample into increasingly more homogeneous subgroups. Fitted values of the outcome variable are estimated for all of the terminal nodes of a tree which will reflect ranges of covariate values in addition to treatment status. Given a regularization procedure to prevent model over-fit (Mullainathan and Spiess, 2017), the resultant estimates prove useful for estimating treatment heterogeneity across a vector of treatment assignments and covariates.

BART employs an MCMC simulation strategy for generating individual estimated outcomes given the covariate values of interest. For a set of N observations, and a $N \times K$ vector of covariates including the treatment variable, BART generates a posterior draw of 1000 predicted values for each unique treatment and covariate profile after a model burn-in phase (Green and Kern, 2012). We treat the average of each of these 1000 draws as the estimated outcome *given* the observed treatment value and covariate profile.

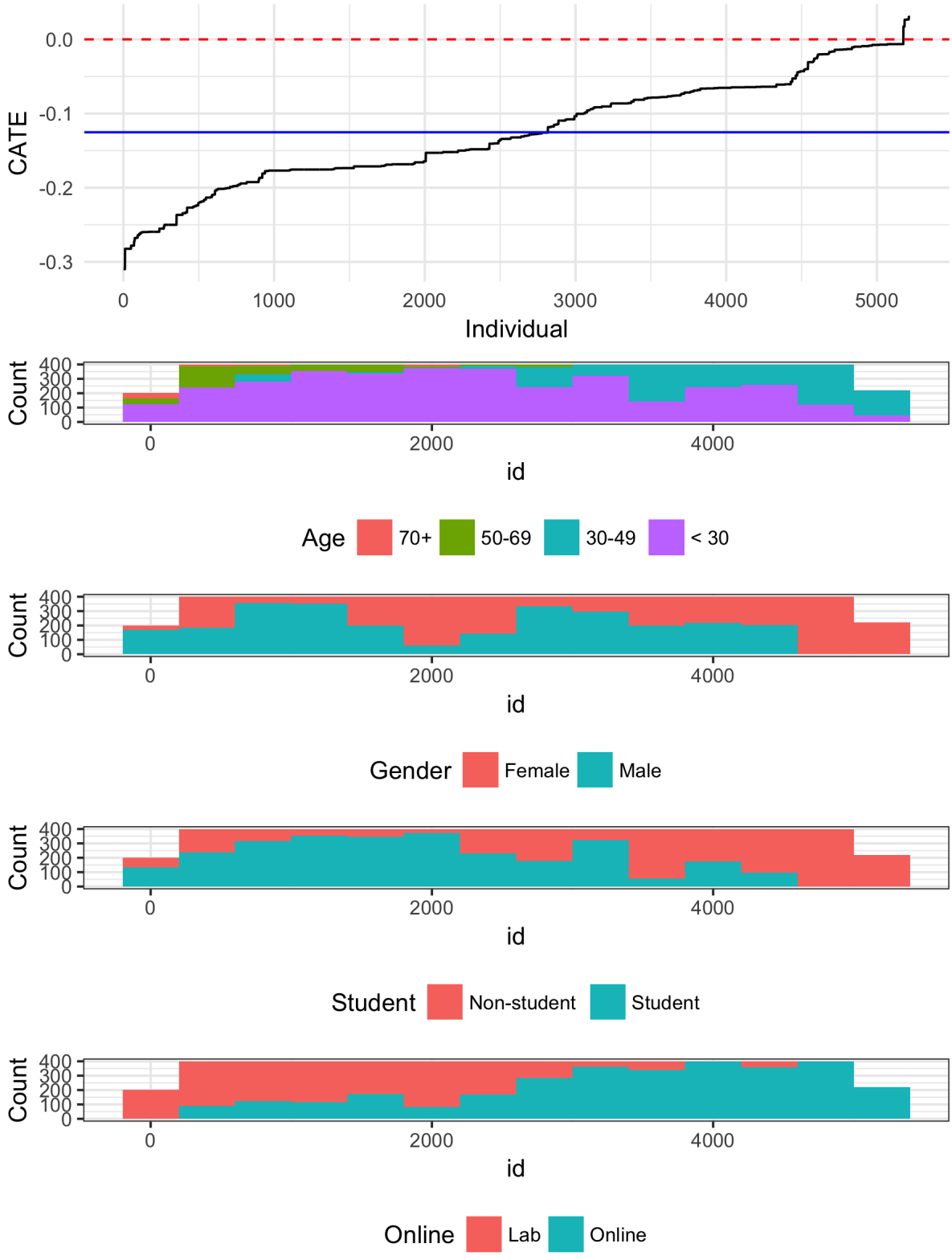
Since the implemented BART procedure predicts outcomes rather than coefficients, we recover CATE estimates by first simulating outcomes for the observed data, and then for a set of counterfactual observations. For the first set of simulated outcomes the BART model takes as inputs the outcome variable of the study (in our case cheating levels) and a training data matrix consisting of the actual treatment assignments and covariates of interest. The second set of simulated outcomes is based on a separate test data matrix. This dataset contains "synthetic" observations that are identical to the training data, except that the treatment

assignments are reversed. This test dataset does not influence the estimation procedure itself. Rather, these counterfactual cases are used post-estimation to predict counterfactual outcomes given the results of the BART model using the observed, training data. Estimating outcomes for *both* the observed and synthetic observations ensures that for any unique set of covariates that have treated cases there will be a matched set of counterfactual "control" cases at that set of values, and vice versa. The CATE for the various covariate values is simply the difference between the *predicted* outcome for the covariate value in the training dataset and the matched observation in the synthetic, test dataset.

We estimate the BART CATEs for the same set of covariates and experimental mode variables that were estimated above using FindIt. Our BART model of heterogeneous effects is a simple specification generated using the BayesTree R package with inputs described as above. All other options within the BayesTree package are left at their default value.

The distribution of CATEs organized by magnitude along with the histogram of covariate and mode/sample pool profiles are presented in Figure 6. The distribution tracks very closely the CATEs estimated with FindIt. The overall average ATE is -0.125 and the range over all the covariate values is -0.31 to just above 0. Clearly, the treatment has a negative effect but there is quite distinctive heterogeneity.

Figure 6: BART estimated heterogeneous effects including covariate interactions



The histograms organized in the lower part of Figure 6 provide again a sense of what experimental conditions are most likely to influence the magnitude of treatment effects. The results are strikingly similar to Figure 5. Almost all of the subjects who played the game in the lab are concentrated at the negative extreme of the distribution of CATEs which is precisely what we saw in Figure 5. Again, consistent with Figure 5, subjects who played the game on-line clearly are distributed over a much broader range of CATEs.

In total we observe over 5,000 decisions in four identical experiment conducted with different modes and subject pools. We estimate the impact of their ability on lying. Automated iterative statistical estimators allow us to identify whether any particular covariates, including our four experimental modes, are responsible for heterogeneity in treatment effects. Two different such estimations of heterogeneity effects result in very similar conclusions: the lab mode results in higher treatment effects while subject pool characteristics do not seem to matter much.

Second, the distribution of CATEs generated by both FindIt and BART suggest that virtually all of the CATEs are negative which is consistent with the initial conjecture and with the estimated ATE in Table 3. And at least in the FindIt estimation, over 80 percent of the CATEs were less than -0.1 suggesting that to the extent that there is subject heterogeneity it tends to be consistent with the overall direction of the overall ATE.

The lab mode generates larger estimated treatment effects for this particular interactive incentivized experiments. There are a variety of possible explanations. We favor a particular underlying mechanism here – compliance to the intended treatment is considerably higher for lab subjects. This can be related to a variety of factors – the manner in which instructions for incentivized, interactive experiments are communicated in the lab, the credibility of incentives, the absence of distractions, etc. But of course one could propose less favorable explanations such as the clarity of the experiment demand effect.

We are inclined to read this evidence as indicating there are features of the lab experiment that enhance compliance to intended treatments. At the same time though our

evidence suggests that the on-line mode does not exaggerate non-compliance in interactive experiments to the point of contracting qualitative results reported in the lab. We want the power and diversity offered by on-line experimental modes. Hence, the challenge is to identify strategies that allow scholars to conduct experiments on-line while at the same time retaining key features of the lab mode that ensure subjects are in fact complying with the intended treatments.

5 Discussion

Concern with fragile treatment effects and replicability of results has lead many scholars to implement their experiments with larger and more diverse subject pools. But of course as we scale-up and diversify, particularly on-line, it becomes more difficult to control the administration of treatments. And as we explore more diverse sample pools, there is a concern that their composition can potentially bias treatment effects. A key issue for many scholars has been simply understanding how, or even whether, treatment effects are affected by experimental modes (Levitt and List, 2015).

This essay reports on a unique experiment that identifies the mode and subject pool effects associated with an interactive experiment conducted in four different experimental conditions. It is unique in that identical incentivized interactive experiments are implemented on-line and in the lab with different types of subject pools. We examine three broad types of measures associated with experiments: 1) the distribution of basic covariates; 2) the choices made in decision-theoretic incentivized games; and 3) decisions made in incentivized interactive games.

First, basic covariates differ of course by convenience sample. Student samples, for example, are much younger than an on-line convenience sample. MTurk subjects are not representative of the general population. There will of course be experiments for which subject pool characteristics are the source of heterogeneous treatment effects (Huff and Tingley,

2015). But in the case of our lying experiment subject pool characteristics per se had only small effects on treatment outcomes.

Second, with respect to choices in decision-theoretic incentivized games, we do not find dramatic differences in effects. We do find some evidence of heterogeneity associated with experimental modes in other-regarding preferences as measured by the Dictator Game. The differences are weak. Nevertheless, they hint at a mode effect – subjects in the lab tend to be more homo economicus than subjects on-line. Most telling is the fact that Oxford student subjects gave more in a Dictator Game if they played on-line than if they played the game in the lab. We can only speculate as to why. One might expect, a priori, that experimenter effects in the lab would result in more generous giving by lab participants. But that is not the case.

Third, we observe choices and treatment effects in incentivized real-time interactive experiments in which subjects have opportunities to lie. Here there is evidence of mode effects. Student subjects playing the game in the lab (as opposed to student subjects playing the game on-line) are more comfortable about lying. Amongst on-line subjects, MTurk subjects exhibit lower levels of lying behavior. MTurk workers belong to a community of crowd-sourced workers (many communicate actively with each other on various forums) and MTurk workers are explicitly evaluated by experimenters. Both of these factors might make MTurk workers less enthusiastic about lying particularly when their lies affect the earnings of other MTurk workers.

A key finding is that ability is correlated with lying. To assess the heterogeneity of this treatment effect across these four experimental modes we implemented two automated non-parametric estimation methods: FindIt and BART. They produce a full set of non-linear interactions with the treatment effect. This facilitates the assessment of heterogeneity across experimental modes and hence of the robustness of the ATE.

For this experimental protocol, these estimation strategies suggest two conclusions: Treatment effects seem to be, at best, weakly correlated with the characteristics per se of the

convenience samples employed. At least with respect to the experiments implemented in this project, there is much stronger evidence of a mode effect. Treatment effects estimated from experiments conducted in the lab tend to be larger.

Having a complete distribution of conditional average treatment effects also allows us to assess the robustness of ATE. We did observe heterogeneity associated with a particular mode but the distribution of CATEs suggested there were few covariate interactions with the treatment variable (Ability Rank) that contradicted the overall ATE. Virtually all of the CATEs had the expected negative signs and had absolute values of .1 or greater (recall the overall ATE was -0.17).

Incorporating multiple experimental modes in the design of certain experiments can provide valuable insights into the robustness of treatment effects. Such a design feature is probably not that informative for most decision theoretic experiments in which the subjects' decisions are not affected by, nor are they affecting, in real time, the choices of other subjects. Our results, consistent with much of the literature, suggests that experimental mode and subject pool have relatively little impact on the estimated treatment effects in decision-theoretic experiments.

On the hand, for incentivized interactive experiments, we are more likely to see heterogeneous effects across modes. As a result, incorporating multiple experimental modes into their design can provide useful information about robustness of treatment effects. An agnostic approach to estimating conditional average treatment effects – such as the two we propose in this essay – provides critical information about both the source of any heterogeneity in treatment effects but also whether the distribution of CATEs raises serious questions about the robustness of the estimated average treatment effects.

References

Arechar, Antonio A., Simon Gächter and Lucas Molleman. 2018. "Conducting interactive experiments online." *Experimental Economics* 21(1):99–131.

Athey, Susan and Guido Imbens. 2017. "The Econometrics of Randomized Experiments." *Handbook of Economic Field Experiments* 1:73–140.

Belot, Michele, Raymond Duch and Luis Miller. 2015. "A Comprehensive Comparison of Students and Non-students in Classic Experimental Games." *Journal of Economic Behavior and Organization* 113:26–33.

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.

Briones, Elizabeth M.and Benham, Grant. 2017. "An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples." *Behavior Research Methods* 49(1):320–334.

Broockman, David E., Joshua L. Kalla and Jasjeet S. Sekhon. 2017. "The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs." *Political Analysis* 25(4):435?464.

Camerer, Colin. 2015. *The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List*. Oxford Scholarship Online.

Cameron, A. Colin, Jonah B. Gelbach and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90(3):414–427.

Chang, Linchiat and Jon A. Krosnick. 2009. "National Surveys Via Rdd Telephone Interviewing Versus the InternetComparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73(4):641–678.

Clifford, Scott and Jennifer Jerit. 2014. "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies." *Journal of Experimental Political Science* 1(2):120–131.

Collaboration, Open Science. 2015. "Estimating the reproducibility of psychological science." *Science* 349(6251).

Coppock, Alexander. 2018. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* pp. 1–16.

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, Denise Laroze and Alexei Zakharov. 2018. "Is Cheating a National Pastime? Experimental Evidence." Nuffield Centre for Experimental Social Sciences Working Paper.

Egami, Naoki, Marc Ratkovic and Kosuke Imai. 2018. Package 'FindIt': Finding Heterogeneous Treatment Effects Version 1.1.4. Technical report CRAN.

Esarey, Justin and Andrew Menger. 2018. "Practical and Effective Approaches to Dealing With Clustered Data." *Political Science Research and Methods* p. 1?19.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413?434.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Horton, John J., David G. Rand and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14:399–425.

Huff, Connor and Dustin Tingley. 2015. ""Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research & Politics* 2(3):2053168015604648.

Imai, Kosuke and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign." *Political Analysis* 19(1):1–19.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Programme Evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Kam, Cindy D. and Elizabeth N. Simas. 2010. "Risk Orientations and Policy Frames." *Journal of Politics* 72:381–96.

Levitt, Steven D. and John A List. 2015. *What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?* Oxford Scholarship Online.

Maniadis, Zacharias, Fabio Tufano and John A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104(1):277–90.

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.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31(2):87–106.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2):109–138.

Neumayer, Eric and Thomas Plumper. 2017. *Robustness Tests for Quantitative Research*. Cambridge: Cambridge University Press.

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