Multi-Modes for Detecting Experimental Measurement Error *

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

October 19, 2018

*Nuffield College Centre for Experimental Social Sciences (CESS) Working Paper

Abstract

Experiments should be designed to facilitate the detection of experimental measurement error. To this end, 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. And we propose incorporating standard measurement strategies in the design that help assess whether any observed heterogeneity reflects experimental measurement error. To illustrate our argument, we conduct, and analyze results from, four identical interactive experiments in the lab; online with subjects from the CESS lab subject pool; online with an online subject pool; and online with MTurk workers.

1 Introduction

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). Some of this concern has focused on the appropriateness of different experimental modes (Camerer, 2015; Levitt and List, 2015; Chang and Krosnick, 2009; Coppock, 2018). Should we estimate treatment effects in traditional experimental lab settings; in the field with large scale RCTs; in the field with hybrid lab-in-the-field experiments; or online with very diverse subject pools? To the extent possible, we should do all of the above!

Experiments should be designed to facilitate the detection of experimental measurement error. To this end, 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. And we propose incorporating standard measurement strategies in the design that help assess whether any observed heterogeneity reflects experimental measurement error.

To illustrate our argument, 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 online with 144 subjects from the same CESS lab subject pool. In a third experiment 90 subjects from the CESS UK Online subject pool took decisions in the identical interactive experiment. Finally, 390 MTurk workers, all from the U.S., made choices in an identical interactive experiment.[1]

Experimental Measurement Error. Experimental measurement error occurs when subjects make choices or decisions that are an unintended artifact of the experimental design. As a result we observe an outcome with error. The challenge for researchers is to identify the features of the design that are contributing to this measurement error. By deliberately varying

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

the experimental mode, researchers have an opportunity to observe heterogeneous treatment effects that might be the product of measurement error. And by embedding metrics within the protocol design, researchers can better assess whether any observed heterogeneity is associated with experimental measurement error.

Heterogeneity. We can think of multi-mode experimental design as essentially blocking on potential sources of experimental measurement error. Within each "block", or experimental mode, there is random assignment to identical treatments and control. What is the point? The debate in the literature concerning experimental modes is essentially a debate about experimental measurement error. Proponents, or critics, of any particular mode typically claim that there are design features that either minimize, or exaggerate, experimental measurement error. Our expectation is that experimental measurement error will likely be more or less prominent in particular modes depending on the design. So, yes, for example, it is likely to be the case that experiments in which subjects are asked to make decisions about "sensitive" behaviors will exhibit higher levels of experimental measurement error depending on the experimental mode adopted.

Identifying heterogeneous treatment effects that are simply an artifact of the experimental mode is strong evidence for experimental measurement error. This should at least signal the possibility of fragile treatment effects. Determining whether experimental mode is a source of heterogeneity contributes to establishing the robustness of an hypothesized treatment effect (Neumayer and Plumper, 2017).

We can think of the "null experimental mode hypothesis" as the assertion that the chosen experimental mode, relative to other experimental modes, exhibits no significant experimental measurement error. And most experimental designs essentially, a priori, assume the "null experimental mode hypothesis" is true (or, presumably, the researcher would have selected a different mode). Experimental modes should be treated like co-variates that could potentially condition treatment effects. By incorporating different experimental modes in

the design and "blocking" on them when assigning subjects to treatments and control, researchers can garner some evidence regarding the null experimental mode hypothesis.

To determine whether there are significant mode effects we apply iterative machine learning methods designed for estimating heterogeneous treatment effects (Athey and Imbens, 2017; Grimmer, Messing and Westwood, 2017). The estimation is conducted without any a priori specification of the functional form of the heterogeneity in treatment effects. The method allows us to estimate the magnitude of treatment effects for all possible combinations of relevant co-variates including the "mode" blocks. This estimation strategy combined with an experimental design that employs diverse experimental modes is a useful method for assessing the robustness of treatment effects. To the extent that there is no significant mode-related heterogeneity in treatment effects we gain some confidence that the estimated treatment effects are not confounded with experimental measurement error. Of course, this will only be the case for measurement error that is correlated with experimental mode. If the measurement error is similarly shared (i.e., the same magnitude) across modes then the multi-mode design would be uninformative. But to the extent that there is a correlation between mode and the magnitude of experimental measurement error (an argument frequently made in the literature) then the multi-mode design will be informative. We illustrate with an experimental design that has identical interactive 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.

Measurement The protocols designed for the multi-mode experiments should anticipate measurement error (Loomes, 2005). Experimental measurement error occurs when subjects make choices or decisions that are an unintended artifact of the experimental design. As a result we observe an outcome with error. The typical linear representation of the treatment effect on outcome y_i is:

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i \tag{1}$$

where $\text{Var}(\epsilon_i) = \sigma^2$ and all of the Gauss-Markov assumptions hold. Instead of observing y_i directly, we observe

$$y_{ik} = \delta_k y_i + \theta_k T_{ik} + u_{ik} \tag{2}$$

where k =the number of experimental modes and T is the treatment variable. The parameter values in Equation 2 suggest sources of experimental measurement error.

If all $\delta_k = 1$ and all $\theta_k = 0$ then we only have classic case of random measurement error. A host of factors could be generating this error – unclear instructions from the experimenter, inattentive subjects, credibility of payments of earnings, etc. And while this might generate considerable noise in the decisions made by subjects it is random. Hence we only obtain imprecise estimates of the treatment effects. If some $0 < \delta_k < 1$ and all $\theta_k = 0$ then there is under-reporting on the outcome variable only. As a result, the estimated treatment effect will be biased towards the null. If some $\theta_k > 0$ then we misreport measurement error for a treatment effect. This will result in an inflated estimate of the treatment effect.

Typically, we only observe $k = 1$ and hence have limited information as to whether any of these four conditions hold. But if $k > 1$ and the modes are sufficiently diverse then there is an opportunity to observe variations in measurement error. Our claim is simply that implementing an identical experiment in diverse experimental modes can help identify the presence of experimental measurement error. Why? Very simply, by varying the experimental mode you vary the constellation of contextual factors that are hypothesized to contribute to experimental measurement error, including: effort on the part of subjects, understanding of the decision making task, diversity of the subject pool, experimenter effects, credibility of the financial incentives, and use of decision making aides (the internet), etc. An indication of experimental measurement error is heterogeneity in treatment effects across

modes. This heterogeneity could signal experimental measurement error.

Our expectation is that if there is measurement error then it will likely vary by mode and hence δ_ks and θ_ks will not be identical across different modes or the random error terms, μ_{ik}, will vary by mode. But of course simply observing mode-related heterogeneity is no necessary indication of measurement error – its simply suggestive. By embedding metrics into the experimental design we can observe either directly or indirectly evidence of experimental measurement error. Ideally, these design features should allow us to distinguish how experimental measurement error is affecting the estimated treatment effects. We illustrate two possible sources of experimental measurement error: random and systematic measurement error in the outcome variable.

First we explore whether the outcome variable that is measured with random error. If in Equation 2 the variance in the error term, μ_{ik}, differs signififantly by k then some modes will generate more imprecise, although unbiased, estimates of treatment effects than others. In this case we should observe, for some modes, a much more dispersed set of CATEs than in others. Elements of the experimental design allow us to assess the extent to which decisions in some experimental modes are noisier than in others. We do not provide a comprehensive enumeration of embedded metrics that are informative in this respect. We focus on one – the extent to which subjects appear to make consistent and meaningful choices over the course of the experimental session. A likely source of random experimental measurement error is simply that some subjects are not making meaningful decisions. And this could be correlated with mode-related heterogeneous treatment effects.

Secondly, experimental measurement error may be systematic and result in under-reporting on the outcome variable. A multi-mode design may be instructive here. This would be the case if $0 < \delta_k < 1$, in Equation 2, holds for some modes but not others suggesting that CATEs in some modes are biased towards zero but not in others. Depending on the mechanisms underpinning under-reporting of the outcome variable, we could expect it to vary across experimental modes. Confirming variations in reporting rates on the outcome variable across

modes is relatively straightforward. A very simple indication that under-reporting might affect estimated treatment effects is to observe whether under-reporting on the outcome variable occurs in experimental modes in which the CATEs are biased toward zero.

Illustration. We argue here for the incorporation of two design features that allow researchers to assess the presence of experimental measurement error and hence the robustness of estimated treatment effects: a multi-mode design and incorporation of metrics to calibrate experimental measurement error. As an illustration we describe the design, implementation and analysis of a multi-mode experiment consisting of four modes. The outcome of interest in our principal experiment is lying. A treatment effect of interest is the cost of lying – specifically varying deduction rates. We also argue that lying is correlated with ability. First, we identify heterogeneous treatment effects with automated statistical learning methods. We combine the four experimental results and use BART to estimate the full set of conditional average treatment effects (CATE).[2] Producing the full range of CATEs for all subjects in all four experimental modes is our point of departure for assessing the presence of experimental measurement error. These CATEs are informative but incomplete. Second, we illustrate how embedding certain metrics in the different experiments can facilitate the identification of experimental measurement error.

2 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.

[2]We also report in the Appendix that the results from using FindIt are essentially identical.

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 and Laury (2002) 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 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 among 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. The final module is a questionnaire that measures preferences and socio-demographic characteristics. Variables included are gender, income, ideological self-placement, trust and the Essex Centre for the Study of Integrity test. The details of these games are provided in Duch, Laroze and Zakharov (2018) and in their Online Appendices.

Online Experiment. We also conduct an online 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 online), 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 $1 for Mturk) (compared to the 300 ECUs = $1 in the lab).

In line with the lab treatment, the deduction rates implemented in the online 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.[3]

[3]In the online 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.

3 Results: Lab and Online Experiments

3.1 Sample Covariates

As one would expect, socio-demographics vary across subject pools. The gender distribution of subjects in the lab and online 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, with an unsurprising younger student subject pool in the lab and online. However, in age, MTurk and UK online subjects are not distinguishable at the 95% confidence level (results in Appendix Table A1).

In terms of differences in decision-theoretic preferences, we find that other-regarding preferences are similar across the different subject pools, but there are differences. In the classic Dictator Game (with a 1000 ECUs pie) we find that a large proportion of subjects either allocate nothing or a half of the endowment to the recipients, with an average allocation to recipients of 286 by students in the lab, 303 by students online, 329 by the general UK panel and 307 by Mturk workers. Students appear more likely to offer nothing when they are in the lab, but mode differences are not statistically significant. In contrast, the UK Online panel and Mturk subjects are significantly more generous than the two student subject pools, but are indistinguishable from each other (t-test and Wilcox rank sum tests available in replication material and descriptive statistics in Online Appendix).

In risk preferences elicited through a standard Holt-Laury 2002 instrument, the UK Online subjects are slightly more likely to score 0.4-0.5, within the risk neutral range. However, overall the different subject pools are quite similar and there are no significant differences across modes or samples.

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. In all four experimental modes, subjects were paid to add two randomly generated two-digit numbers in

one minute (payment to online subjects were lower than in the lab). Despite minor variations in the distributions of correct responses, there are no substantive differences in average gains across subject pools or modes (Fig. A4 in the Appendix). The average number of correct responses was 10.13 for UK Online, 10.50 for Mturk workers, 11.06 for students in the lab and 11.85 for students online. The differences are not surprising considering it is an Oxford student sample.

3.2 Treatment and Co-variate 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 report rates will drop as deduction rates rise (a.k.a. higher lying); report rates will be lower when there is no auditing of income; and those who perform better on the RET will lie more about their income.

Table 1 reports results for the regression model with the percent of income reported as the dependent variable. To estimate treatment effects, we include two dummy variables for the 20% and 30% deduction rates, and a "No Audit" dummy variable. The covariate, ability, is 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). In addition, we include age and gender as further controls. The baseline is the 10% deduction rate and a 10 percent audit rate. The Deduction dummy coefficients are negative and significant for the Lab subject pools (but not for the online subject pools). And the Audit dummy variable is negative and significant in three of the four models. For the four online and lab models, the estimated coefficients for Ability Rank are, as expected, negative and significant in all four equations. 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.

	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

Table 1: GLM estimation on percent declared

The estimated effects reported in Table 1 are significant and in the expected direction for the lab experiments; there is more variability in direction and significance for the online multivariate results. In the Appendix Table A2 we report the Wild and PCB p-values for the coefficients reported in Table 1, which further indicate that effects for online experiments are much more imprecisely estimated.

3.3 Heterogeneous Mode Effects?

There is a pattern in the estimated coefficients from Table 1 that suggests variation in treatment effects across modes. Note that the deduction rate treatments are particularly significant, and in the correct direction, for the lab and online lab modes – but weaker and incorrectly signed for Online UK and Mturk. The "No Audit" treatment is quite large,

significant and correctly signed for the Lab experiment but weaker for Online Lab and Online UK and indistinguishable from zero for the MTurk experiment. And the coefficient for ability is strongly negative for the Lab mode but smaller for the other three modes. And demographic co-variates are significant in some, although not all, modes.

Should we conclude the estimated treatment effects are robust across modes? The model estimates reported in Table 1 suggest not. But by using GLM models, and running separate estimations for each mode, we have effectively imposed a particular specification for explaining honest behavior through, essentially, 'ad hoc variable selection' (Imai and Ratkovic, 2013*a*, p.445). While this strategy could be perfectly acceptable in many cases, it is not the optimal strategy for identifying heterogeneous mode effects (which could signal experimental measurement error). At least with respect to estimating possible mode effects typically we have no a priori expectations as to how mode interacts with either the treatment effects or other co-variates. Nor do we necessarily have any priors on how other co-variates interact with treatment variables.

Rather than impose structure on the estimation of the treatment and co-variate effects, and their interactions, we propose a procedure that effectively automates the identification of heterogeneous mode effects: estimating conditional average treatment effects (CATEs) for the combined data from the four identical experiments using iterative, machine-learning procedures. This strategy enables us to probe treatment effect robustness in two ways. First, is there evidence in the CATEs contradicting the conclusions regarding average treatment effects estimated in Table 1? Second, are there significant differences in CATEs across experimental modes?

As we argued earlier, the null hypothesis here is simply that mode-related heterogeneity in treatment effects is not significant. But of course there are lots of other possible sources of heterogeneous effects. The challenge is to identify which interactions of covariates and treatment effects are noteworthy. 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. In the main text we present the results of one automated statistical learning estimation strategy.[4]

A 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 out-

[4]Results for an additional estimation strategy, FindIt, are reported in the Appendix (Table A3 and Figure A5) – the findings are essentially the same as those reported for the BART method in this section.

comes 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 lying) 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.

A CATE is estimated for each subject based on their individual vector of treatment and covariate values, including their mode assignment. 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 BayesTree 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 1. The overall average ATE is -0.06 and the range over all the covariate values is -0.22 to 0.12. The distribution of CATEs gener-

ated by BART suggests that about two-thirds of the CATEs are negative which is consistent with the initial conjecture and with the estimated ATE in Table 1. About half of the CATEs were less than -0.06 suggesting that to the extent that there is subject heterogeneity it tends to be consistent with the direction of the ATE.

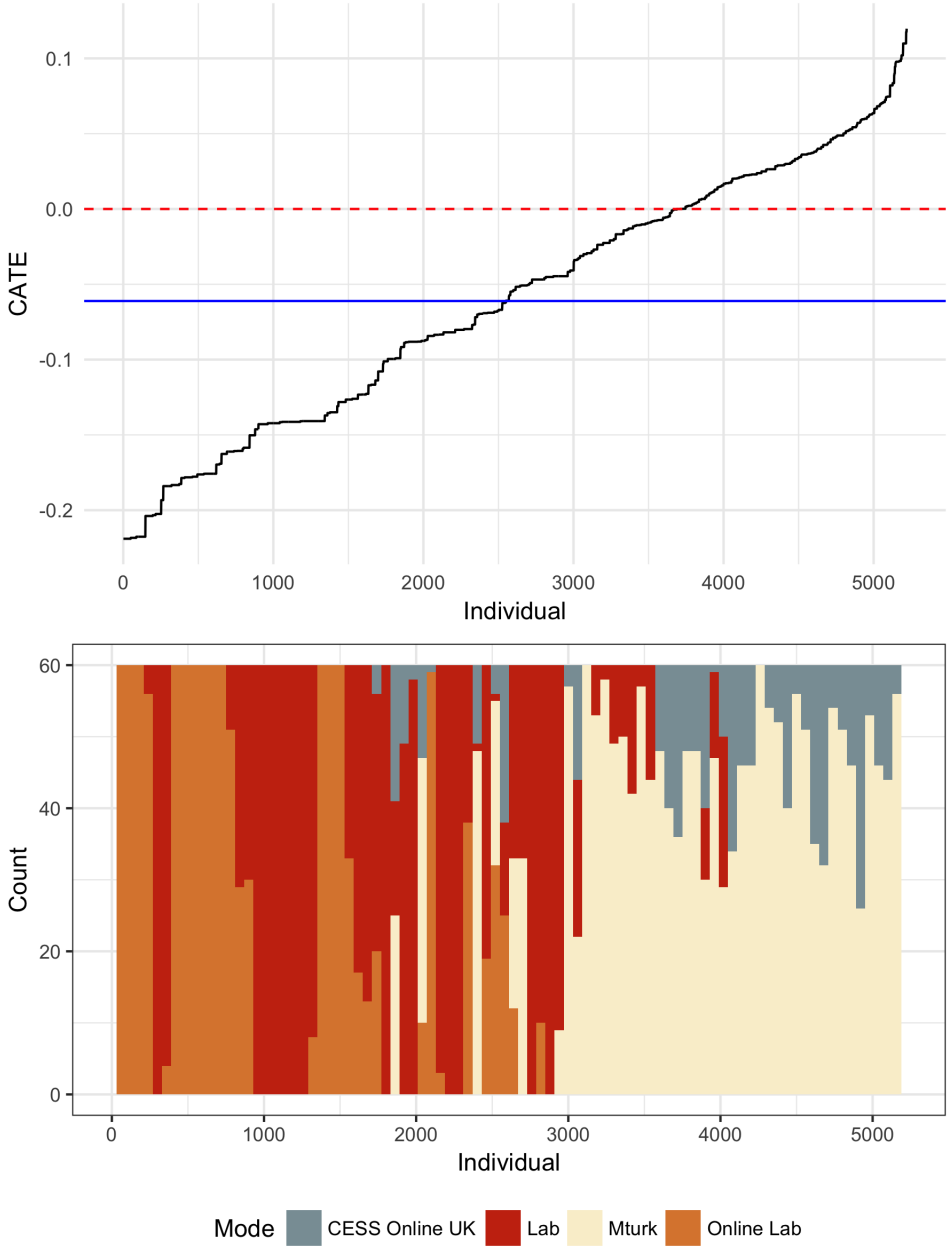
The overall treatment effect is negative but there is distinctive mode-related heterogeneity. The histogram in the lower part of Figure 1 provides a sense of how experimental modes influence the magnitude of treatment effects. Participants from the classic lab subject pool – whether they play the game in the lab or online – exhibit the highest Deduction Treatment effects. Most of these subjects have treatment effects that are more negative than the ATE of -0.06. Online subjects from either MTurk or CESS Online for the most part had CATEs greater than -0.07; and almost a third of these subjects had CATEs that were incorrectly signed.

In total we observe over 5,000 decisions in four identical experiments conducted with different modes. 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, one reported here and the other in the Appendix, result in very similar conclusions: *treatment e ects di er by mode*.

3.4 Measurement Error

The mode-related heterogeneity in CATEs, observed in the previous section, could signal experimental measurement error. There are other possible explanations. Huff and Tingley (2015), for example, suggest that the clustering of particular co-variates in different experimental modes could very well explain mode-related heterogeneity. In the method we propose, though, some of these competing co-variate interactions are accounted for and can be evaluated as competing explanations for mode heterogeneity. Moreover, the experimental protocols incorporated metrics that corroborate whether mode-related heterogeneity reflect

Figure 1: BART estimated heterogeneous effects by mode



experimental measurement error.

Random Measurement Error. Random measurement error in the outcome variable can reduce the precision of estimated treatment effects. A design feature, that helps detect random experimental measurement error is to observe subjects, in different experimental modes, making lots of decisions – either very similar, or identical, decisions or decisions that we expect to be related in a predictable fashion. More generally, there is a growing recognition that an effective strategy for estimating experimental measurement error is to observe subjects making decisions when confronted with similar or identical choice sets (Gillen, Snowberg and Yariv, Forthcoming; Engel and Kirchkamp, 2018).[5]

Our outcome variable in this experiment is the amount of income subjects report after each round of a real effort task (RET). Subjects report in this fashion a minimum of 10 times over the course of an experimental session (20 for students in the Lab).[6] For our assessment of measurement error, we measure the variability of decisions made by subjects within a particular deduction and audit rate treatment. For a particular deduction and audit rate treatment we compare the variability of subjects' decisions over the 10 rounds (intra-subject variability) with its variability across subjects (inter-subject variability). We calculate the Intraclass Correlation Coefficient (ICC) which is simply the ratio of the between-cluster variance to the total variance. It indicates the proportion of the total variance in reported earnings that is accounted for by the subject clustering. We can think of it as the correlation among scores for any particular subject. Our expectation is that between subject variability should account for much of the total variance – hence a high ICC. Moreover, the null hypothesis is not simply that the ICC is high but also that it is very similar across quite different modes.

[5]Random measurement error associated with covariates is particularly problematic because it can result in biased estimates of treatment effects. Strategies for identifying and correcting for this bias agaion build on this practice of observing subjects make multiple decisions, for example on measures of risk aversion (Gillen, Snowberg and Yariv, Forthcoming; Engel and Kirchkamp, 2018). While recognizing that this work is very much complementary to our efforts, we do not specifically deal here with measurement bias in covariates.

[6]Because of a programming mistake in some of the UK Online sessions people only made these decisions 4 times. This was detected quickly and fixed.

Mode	(1)	(2)	(3)	(4)
Lab	0.769	0.905	0.76	0.85
	(0.027)	(0.021)	(0.049)	(0.031)
Lab Online	0.745	0.863	0.633	0.767
	(0.029)	(0.022)	(0.039)	(0.042)
Online UK	0.771	0.92	0.703	0.752
	(0.036)	(0.02)	(0.101)	(0.131)
MTurk	0.808	0.78	0.892	0.828
	(0.022)	(0.016)	(0.028)	(0.03)
Deduction Rate	10%	30%	10%	30%
Audited?	No	No	Yes	Yes

Table 2: Comparison of outcome ICCs across modes

Table 2 presents the ICC for the outcome variable (percent of RET earnings reported) in the experiment. The four columns correspond to different deduction/audit rate treatments, and Table 2 reports ICCs for each of the four experimental modes. Bootstrapped standard errors are shown in brackets. There are no dramatic differences, within any treatment, across modes: in the 10% deduction/zero audit treatment the ICCs range between 0.75 and 0.81; for the 10% deduction/non-zero audit rate the range is 0.63 to 0.89; in the 30% deduction/zero audit they fall between 0.78 and 0.91; and when the treatment is 30% deduction and non-zero audit it ranges between 0.75 to 0.85. The one potential outlier here is the 0.63 ICC estimated for the Lab Online mode.

Subjects in this experiment, at least with respect to the outcome variable, appear to behave quite consistently across many rounds of identical decision making tasks. And consistent behavior is observed across quite different experimental modes. There is little evidence at least for this one metric to suggest that random measurement error is correlated with experimental mode.

An additional metric of consistent behavior is the subjects' performance on the real effort task (RET). Again we leverage the fact that subjects perform these real effort tasks either 10 or 20 times over the course of an experimental session. We compare the consistency

Mode	(1)	(2)	(3)	(4)
Lab	0.768	0.768	0.636	0.85
	(0.018)	(0.018)	(0.039)	(0.047)
Lab Online	0.807	0.76	0.762	0.767
	(0.018)	(0.017)	(0.021)	(0.047)
Online UK	0.88	0.827	0.827	0.752
	(0.011)	(0.018)	(0.026)	(0.029)
MTurk	0.758	0.758	0.782	0.828
	(0.015)	(0.012)	(0.024)	(0.026)
Deduction Rate	10%	30%	10%	30%
Audited?	No	No	Yes	Yes

Table 3: Comparison of RET ICCs across modes

of subject performance across the four experimental modes. We employ the same strategy adopted for earlier for the earnings reporting variable. Table 3 reports the ICC calculated for each of the four modes controlling for treatments. We observe quite high ICC values for the CESS Online UK mode suggesting these subjects were particularly consistent in their RET performance across rounds. But overall the ICCs were quite high, and consistently so, across treatments for subjects in all four modes. There is no strong evidence here of measurement error; nor evidence that it varies significantly across modes.

A third metric for estimating random measurement error is to assess the extent to which subjects respond in a similar or consistent fashion to items that have been demonstrated to measure an underlying attitude. Again, the expectation is that inconsistent responses to such items would signal random measurement error. And to the extent that we observe variation in this inconsistency across modes we might conclude that there is in fact experimental measurement error. The experiment included a series of questions that make up the Essex Centre for the Study of Integrity (ECSI) test (Whiteley, 2012). They have been administered widely and the items are highly correlated. Table 4 reports the Cronbach Alpha scores for subjects in the four experimental modes. Consistent with our other measures, subjects answer in a consistent fashion: The Cronbach Alpha coefficient is typically around the acceptable 0.7 level; and this consistency is observed at similar levels across all four modes.

Only Lab students online are slightly less consistent.

Mode	Cronbach's Alpha	95% Lower Bound	95% Upper Bound
Lab	0.712	0.616	0.771
Lab Online	0.636	0.483	0.721
CESS Online	0.715	0.575	0.801
MTurk	0.749	0.694	0.792

Table 4: LTM Cronbach's alphas for integrity responses

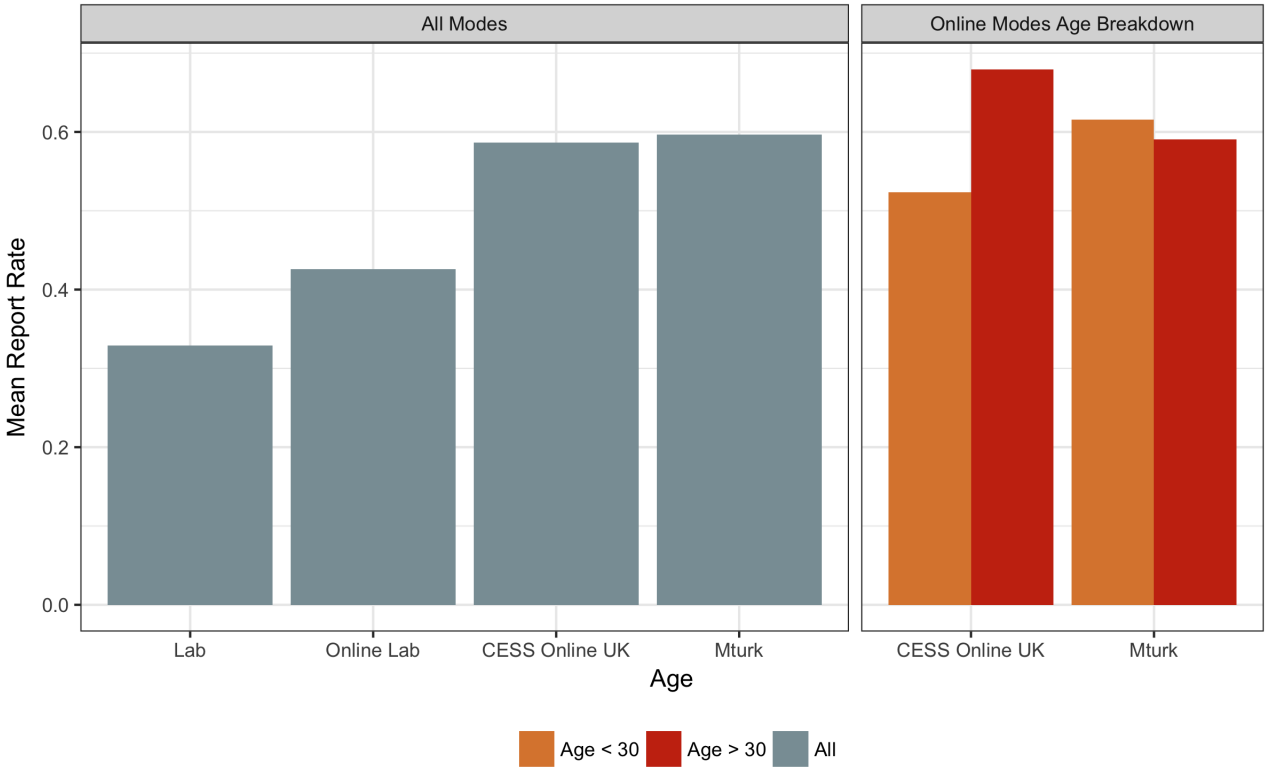
Systematic Measurement Error. As we pointed out earlier, a source of systematic measurement error is under-reporting preferences or behaviors measured by the outcome variable. Typically this occurs when subjects under-report sensitive behaviors or decisions. This biases treatment effects toward the null. Incorporating diverse experimental modes in the design can facilitate the identification of this systematic measurement error.

The challenge is incorporating features into the design that convincingly identify whether measurement error is generated by under-reporting (Blattman et al., 2016). Lying about earnings, the outcome variable in our experiment, is plausibly a sensitive choice for subjects to make. And there is an extensive measurement literature on how reporting of sensitive behavior varies by experimental, or survey, mode (Tourangeau and Yan, 2007; Tourangeau, Rips and Rasinski, 2000). We assume here that diverse experimental modes trigger varying concerns regarding the social desirability of certain reported behavior. There clearly is evidence that treatment effects in our experiments are not significant in some modes – possibly the result of under-reporting.

The mode-related heterogeneity in CATEs observed in Figure 1 indicated that subjects in the MTurk and CESS Online modes had CATEs closer to zero or, for many, incorrectly signed. This could result because these participants from online subject pools are hesitant to lie about their earnings. We are able to compare the rates of lying across experimental modes that can provide some insight into whether under-reporting might be a source of

measurement error. The left-hand graph in Figure 2 reports the incidence and magnitude of lying across the four modes.

Figure 2: Comparison of Income Report Rate



There are two behavioral differences that stand out for the zero-audit condition in Figure 2. First, subjects, for the most part Oxford undergraduate, drawn from the lab subject pool (Lab and Online Lab) are more comfortable lying about their income. Second, subjects from the online subject pools (CESS Online and MTurk) are more hesitant about lying. In the zero audit condition, lab subjects overall report between 30 and 40 percent of their earnings while online subjects report about 60 percent of their gains. These results are consistent with the notion that under-reporting on the outcome variable (lying) contributes to the null findings observed in Figure 1 for the MTurk and CESS Online modes.

An alternative explanation, of course, is that age and mode are confounding variables in these comparisons. The observed higher levels of earnings reported for CESS Online

and MTurk modes might simply reflect the presence of older subjects in the sample (the lab samples were students and hence essentially young). The right-hand graph in Figure 2 indicates this may not be the case. Here we control for the subjects over and under 30 years of age for the zero-audit condition. For the CESS Online mode there is some evidence here that older subjects drive some of the underreporting. But for the MTurk modes the two age cohorts have essentially identical levels of reporting (or lying).

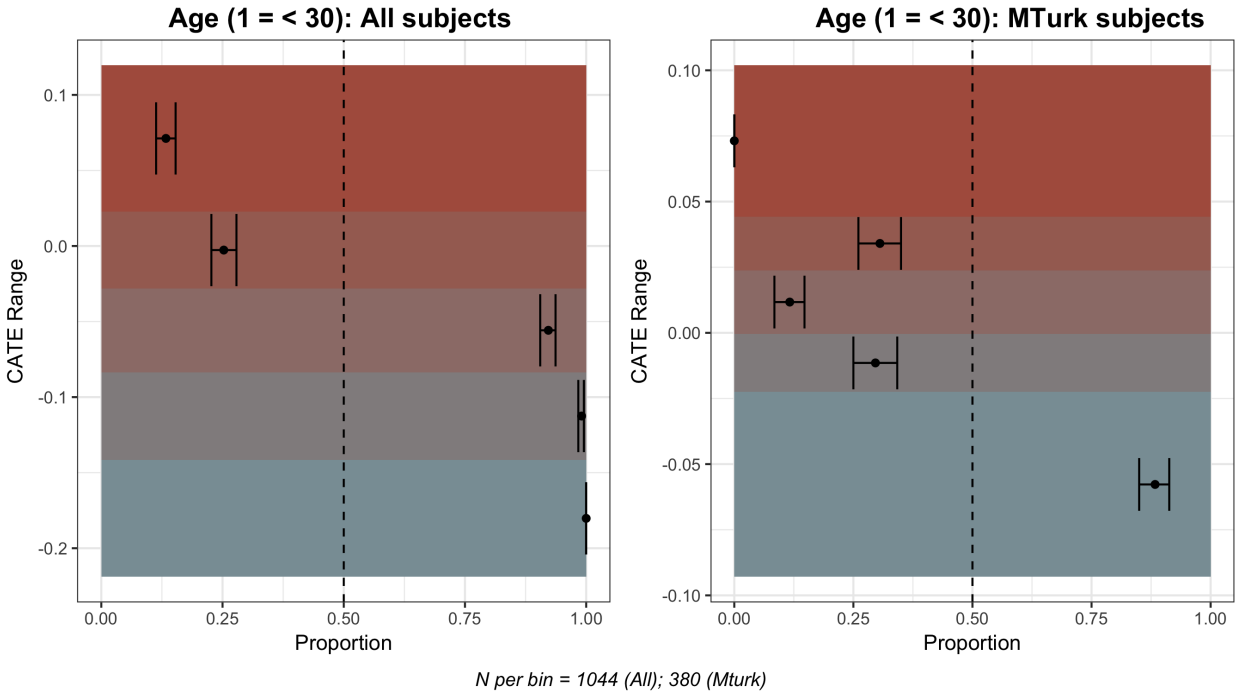


Figure 3: Banded bootstrap test for age covariate - all subjects versus MTurk only

There is some evidence here that subjects from online subject pools, MTurk in particular, are reluctant to lie about their earnings; and, at least in the MTurk case, this does not seem to be related to the age differences between online and lab subject pools. As Equation 2 suggested, under-reporting (lying in this case) can bias the estimated treatment effect to the null. And this would seem to be the case here for both older and younger MTurk subjects. Figure 3 compares the distribution of age cohorts across different bands of CATEs for all four modes combined and also for the MTurk mode. The estimated treatment effect is considerably lower in the case of the MTurk mode. The highest treatment effect band for

the MTurk mode straddles -0.05 where we do see a disproportionate concentration of young subjects. But in the case of the combined four modes we see treatment bands extending to -0.20; and in this band we also see a high concentration of young subjects.

Subjects from the CESS lab subject pool, regardless of whether they played the game online or in the lab, were significantly more likely to lie than was the case for CESS Online or MTurk subjects. Moreover, this difference persists for the MTurk mode even when we control for age, i.e., comparing the lab subject pool (who are essentially all young students) with young subjects from the MTurk subject pools (younger subjects who are not necessarily students). As a result, the treatment effects in Figure 1 are much larger for subjects from the Oxford lab subject pool, regardless of whether they played in the lab or online.

4 Discussion

A classic experimental design is one in which subjects are randomly assigned to treatments and make incentivized decisions that are influenced by, and influence, the choices of other subjects in real time. Most of these experiments take place in a classic experimental lab setting with convenience samples, often of students. A concern is the fragility and replicability of treatment effects generated in these experimental settings. This essay reports on experimental designs that provide insights into the robustness of treatment effects estimated for incentivized interactive experiments.

Most importantly, the design should incorporate scale and diversity. Scale goes without saying – there is power in numbers. It is now the case that experiments can be conducted in very diverse contexts – ones that can vary quite significantly in terms of subject pools and modes. If replicability matters then relying on the estimated treatment effects from a particular experimental mode and subject pool is risky. Hence, scholars should have a good sense of how their estimated treatment effects vary across diverse experimental modes. As an illustration, this essay reports on a unique experiment interactive experiment conducted

in four different experimental modes. It is unique in that identical incentivized interactive experiments are implemented online and in the lab with diverse subject pools.

The multi-mode feature of the design speaks to the robustness of estimated treatment effects. The modes here represent distinct settings in which identical experiment designs are implemented. Our assumption here is that experimental measurement error will vary across these modes. We think of mode as a co-variate in a block design. Subjects are randomly assigned to treatments within each of our four "blocks". The null hypothesis is no significant interaction between this co-variate and the treatment effects. Of course there are a number of competing co-variate interactions that might be affecting the treatment effects. This essay suggests a strategy for estimating the mode interaction effect along with any other plausible co-variate interactions.

We adopt a non-parametric iterative estimation techniques to assess whether the four experimental modes condition estimated treatment effects. This generates CATEs for all subjects who share particular values on all combinations of relevant co-variates including the experimental modes in which they participated. Assuming mode is not conditioning, or interacting with, treatment effects, the CATEs for subjects from the four different experimental modes should be similarly distributed over the full range of CATE values. There is noteworthy mode-related heterogeneity in our treatment effects. Most notably, the treatment effect is much stronger for subjects recruited from the Oxford student lab subject pool, regardless of whether they play the game online or in the lab. On the other hand the treatment effect is quite weak, although overall correctly signed, for the MTurk and CESS Online modes.

Mode-related heterogeneity suggests the presence of experimental measurement error – but it is only suggestive. An alternative explanation might simply be the composition of the convenience sample. Experimental protocols should incorporate design elements that can help confirm the presence, or absence, of experimental measurement error but also characterize the nature or source of this measurement error. We should think of experiments

as measurement instruments that require calibration depending on how they are being implemented. Figuring out the character of experimental measurement error associated with different modes helps determine how the experimental instrument needs to be calibrated from one mode to the next. For example, what should we conclude if we conduct identical experiments in the lab and online, observe the same estimated treatment effect, but with much more uncertainty online? This might simply be a calibration problem – the experiment or "instrument" measures ATEs more imprecisely online. Or, the online results might suggest a real null effect.

We suspect, as do many others, that the types of experimental measurement error likely to be observed may differ by experimental mode. Hence, implementing identical experiments in different modes is unlikely to generate identical treatment effects, regardless of what the underlying reality might be. But having a good knowledge of how experimental measurement error varies by mode helps to appropriately "calibrate" the experimental instrument or the estimated treatment effects. Broadly speaking we distinguish between random and systematic measurement error that can affect the observed values of both outcome variables and co-variates when we conduct an experiment. Some version of this experimental measurement error could cause mode-related heterogeneity in estimated treatment effects. We illustrate strategies for identifying experimental measurement error in our multi-mode experiments on lying.

Random experimental measurement error in the case of the outcome variable can reduce precision and in the case of covariates can bias estimated treatment effects. As part of our experimental design we had subjects, controlling for treatments, performing identical tasks and also making identical decisions on multiple occasions over the course of the experimental session. We treat the consistency of these decisions within subject, compared to inter-subject variability, as one indicator of random measurement error generated by the experimental design. Subjects performed the same real effort tasks (one-minute for addition of two two-digit numbers) ten times; and subjects reported their total earnings (with an opportunity to

lie) on ten different occasions. The results for the four different implementations of our lying experiment were re-assuring: our measure of subject consistency, Intra-Class Correlation Coefficient, was generally quite high; moreover, it was similar across all four experimental modes.

In all four implementations of the experiment we also asked subjects to answer a battery of items designed to measure Integrity. There is evidence to suggest that answers to these questions should be highly correlated. We use estimations inter-item correlation across the four experimental modes as another indication of experimental measurement error. The results are reassuring: the Cronbach Alpha statistic was consistently high across all four modes.

We also explored whether systematic experimental measurement error played any role in the observed mode-related heterogeneity. Specifically, we identified systematic experimental measurement error generated by under-reporting of choices perceived to be socially undesirable. The outcome variable in this experiment is lying – misreporting of earnings in a real effort task. Under-reporting on the outcome variable biases treatment effects towards the null. Again, its difficult to demonstrate this might condition an estimated treatment effect unless we can observe some variance in possible under-reporting. In our illustration we do in fact observe under-reporting of lying for the CESS Online and MTurk experimental modes. And as the measurement model predicts we find tha the CATEs for the MTurk and CESS Online modes are concentrated in the zone of a null treatment effect.

Short of outright fraud, experiments do not replicate either because of knife-edge treatment effects (Gelman, 2013) or experimental measurement error. This essay focuses on the latter. A single implementation of an experimental protocol either in the lab, online or in the field is unlikely to identify experimental measurement error. The researcher in this case has little information about the robustness or the replicability of the estimated treatment effects. Implementing quite distinct modes of an experiment design increases the likelihood that the researcher observes experimental measurement error. This essays suggests strate-

gies for extracting meaningful measurement error information from the resulting data. And we strongly endorse Gillen, Snowberg and Yariv (Forthcoming) who argue for incorporating metrics in the design that allow for the identification of experimental measurement error. We suggest that these metrics along with the multi-mode design can provide an indication of the presence of experimental measurement error and how it might affect estimated treatment effects.

References

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.

Blattman, Christopher, Julian Jamison, Tricia Koroknay-Palicz, Katherine Rodrigues and Margaret Sheridan. 2016. "Measuring the measurement error: A method to qualitatively validate survey data." *Journal of Development Economics* 120:99 – 112.

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.

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.

Duch, Raymond, Denise Laroze and Alexei Zakharov. 2018. "Once a Liar Always a Liar?" 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.

Engel, Christoph and Oliver Kirchkamp. 2018. "Measurement Errors of Risk Aversion and How to Correct Them." Working Paper.

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

Gelman, Andrew. 2013. "Preregistration of studies and mock reports." *Political Analysis* 21(1):40–41.

Gillen, Ben, Erik Snowberg and Leeat Yariv. Forthcoming. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* .

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.

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

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*a*. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.
URL: *http://projecteuclid.org/euclid.aoas/1365527206*

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

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

Loomes, Graham. 2005. "Modelling the Stochastic Component of Behaviour in Experiments: Some Issues for the Interpretation of Data." *Experimental Economics* 8(4):301–323.

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.

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

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

Tourangeau, Roger, Lance Rips and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 5:859–83.

Whiteley, Paul. 2012. "Are Britons Getting More Dishonest?" ECSI Working Paper.

Appendix A Descriptive statistics

A1 Experimental Sessions

Table A1 presents a summary of the experiments and treatments (audit rates and deduction rates) incorporated in each of the experimental modes, the number of subjects that participated, the percentage of female to male subjects, as well as the mean report rate. AS can be observed, there gender distribution is relatively balanced, except for the Lab mode where 44% of participants are female. Complementary research conducted by the authors suggests this is not a problem as there are no male/female differences in lying in this game. Audit rates were either 0, 0.1 or 0.2, all online version included 0 and 0.1 audit rates, and lab included 0.2, this slight variation is not expected to generate any issues, as report rates are lower in the Lab, despite a higher probability of being audited. The deduction was 10 and 30% for all online modes, while in the lab there were also sessions with 20% deduction. Subjects in all modes completed a Dictator Game and a risk aversion lottery.

Mode	DG	Risk	Audit rate	Deduction rate	Report rate	# Subjects	% Female	% Male
CESS Online UK	Yes	Yes	c(0, 0.1)	c(10, 30)	0.63	90	0.52	0.48
Lab	Yes	Yes	c(0, 0.2)	c(10, 20, 30)	0.43	116	0.44	0.56
Mturk	Yes	Yes	c(0, 0.1)	c(10, 30)	0.60	390	0.49	0.51
Online Lab	Yes	Yes	c(0, 0.1)	c(10, 30)	0.46	144	0.50	0.50
All	Yes	Yes	c(0, 0.1, 0.2)	c(10, 20, 30)	0.53	740	0.48	0.52

Table A1: Summary of experimental treatments

A2 Sample Covariates

Socio-demographics vary across subject pools. AS indicated in Figure A1 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 online are even younger on average. The general UK online panel subjects are similar to MTurk subjects. The age

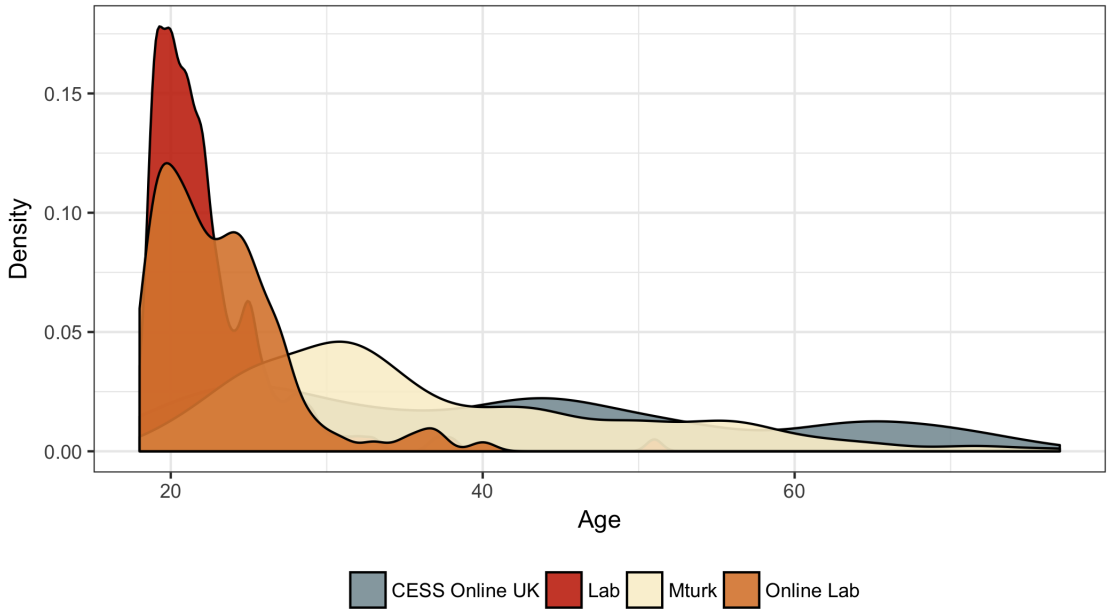


Figure A1: Age distribution of subjects

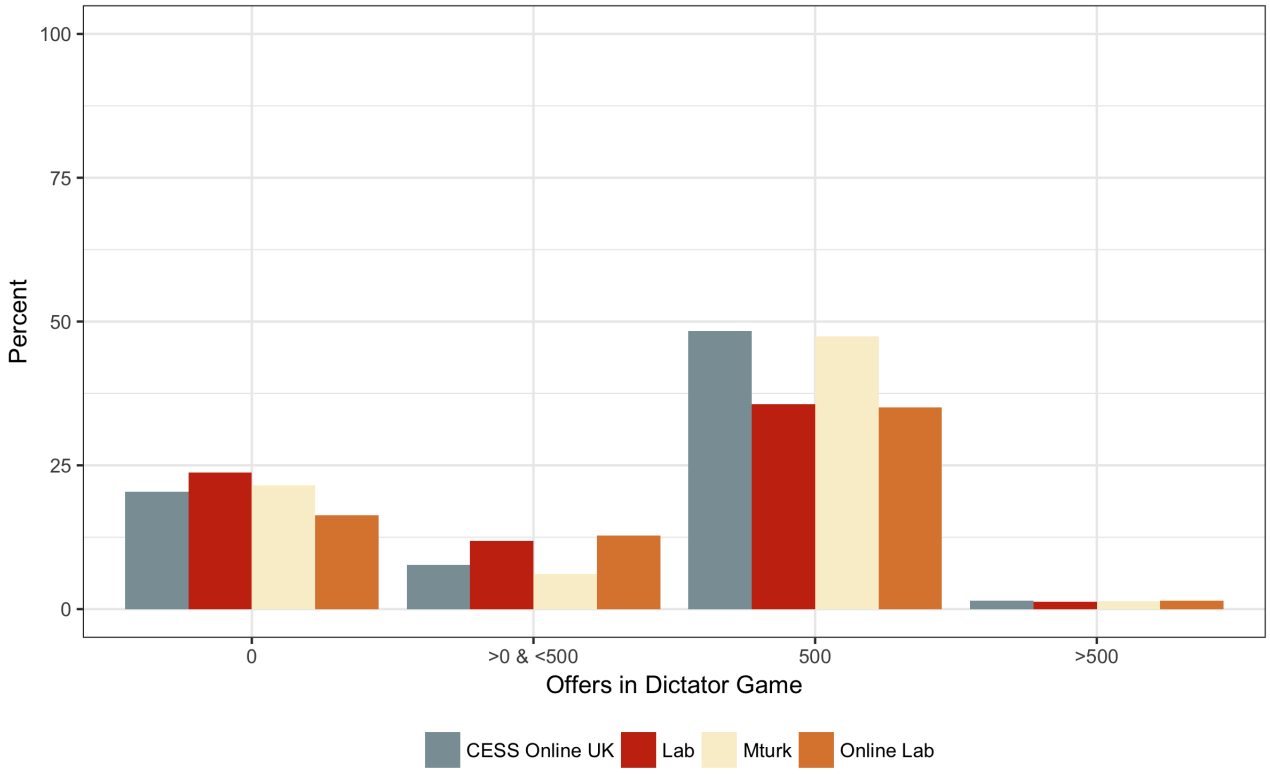
distributions for MTurk and UK online are significantly different from UK lab and online in both t-test and Wilcoxon rank sum test, but MTurk and UK online are not distinguishable at the 95% confidence level.

A3 Decision-theoretic preferences

One concern is that subject pools may differ with respect to fundamental preferences Belot, Duch and Miller (2015). We implemented a set of incentivized decision theoretic experiments designed to recover a number of standard preferences.

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 online versions of the Dictator Game subjects have an opportunity to split an endowment of 1000 ECUs between themselves and an undisclosed recipient. Figure A2 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

Figure A2: Dictator Game



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 online, 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 online is insignificant. In contrast, the UK Online 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 UK online panel are indistinguishable from each other.

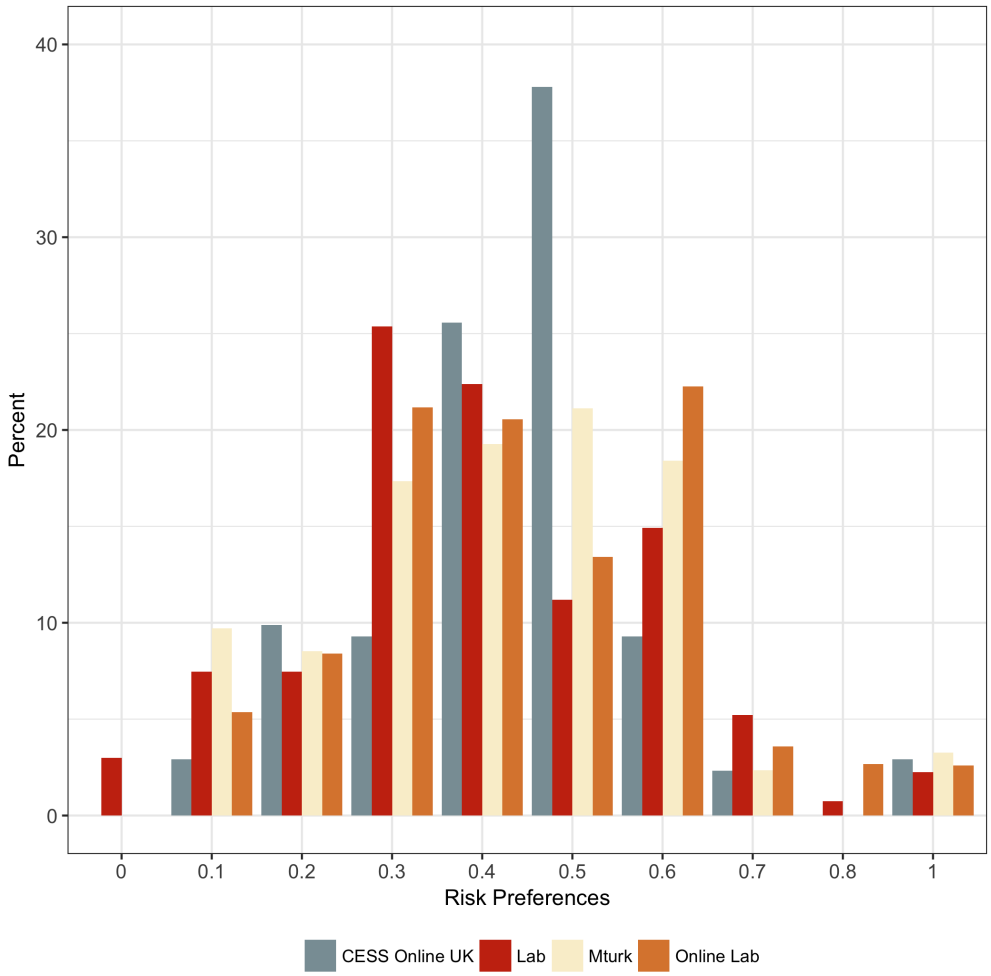
A second incentivized experiment elicited the risk preference of both lab and online subjects employing a standard Holt-Laury (2002) instrument. Participants were asked to make ten choices between two lottery's Option A (less risky) and Option B (more risky)

– screen-shot in replication material. 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 online, 16 percent of students in the lab, 17 percent of Mturk workers, and, a surprisingly high, 31 percent of CESS Online subjects. Eliminating these observations from the analyses does not substantively alter the results, therefore observations are kept to avoid reducing the sample size.

Figure A3 shows the distribution of risk preference from the studies. The x-axis in Figure A3 presents a ratio of the number of times a participant chose Option B over the total ten decisions. CESS Online 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 online 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 online experiments, a risk preference of zero could result from 1) the participant logging off (in those cases the code recorded the answers as zero/Option A); or 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 online (Cess online UK: $p < 0.001$; MTurk: $p < 0.05$) although the difference between subjects from the same lab subject pool playing the game online and in the lab does not reach conventional levels of significance ($p > 0.1$). And at least two of the three different online subject pools

Figure A3: Risk Preference



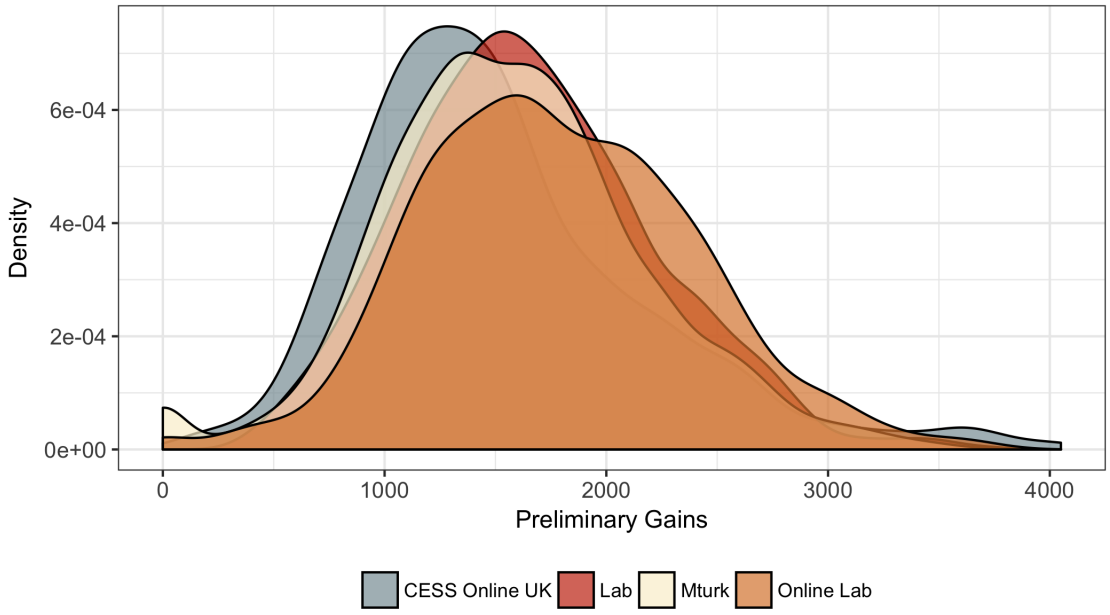
made very similar average offers in the Dictator Game. On the second incentivized risk preference experiment subjects made similar choices – none of the risks results from the four experiments suggested a significant mode or sample difference.

A4 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 online subjects were lower than in the lab). Figure A4 shows the distribution of outcomes for both lab and online 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 Online 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) online. Student subjects (Lab and Online) are primarily Oxford undergraduates and are, on average, better educated which might explain the higher performance. MTurk subjects performance is slightly higher than UK online, possibly a result of being "professional" online workers.

Figure A4: Real Effort Task Performance



A5 Robustness tests on estimations

Table A2: 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

As a robustness test for standard errors presented in table 1 – that could potentially understate the uncertainty for the online experiments due to the small number of subjects (Esarey and Menger, 2018) – we estimated GLM models using wild cluster bootstrapped t-statistics ("Wild") and pairs-clustered bootstrapped t-statistics ("PCB") respectively (Cameron, Gelbach and Miller, 2008). The results, presented in Table A2, indicate the significance of our coefficient estimates diminish substantially, as expected. However, 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 online lab participants. Deduction rates of 30 % also remain significant ($p < 0.05$) for online lab participants.

As another robustness test we follow the Support Vector Classifier (SVC) suggested by Imai and Ratkovic (2013*b*). The iterated LASSO model estimates produced by the Imai and

Ratkovic (2013*b*) 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 – online 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 online dummy variables, as well as age and gender covariates (also included in the interaction).[7] In line with Imai and Ratkovic (2013*b*), 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*b*) provide a detailed discussion and full specification for the GCV statistic used.

Table A3: Heterogeneous treatment coefficients and interactions using iterated LASSO model

Variable	Coefficient
Treatment	-0.051
online Lab	0.014
MTurk	0.054
CESS online UK	0.079
Age	0.004
Gender	0.012
Treatment × MTurk	0.067
Treatment × CESS online UK	0.189
Treatment × Ability	-0.011
Treatment × Gender	0.041
Online Lab × Ability	0.060
Online Lab × Age	0.0003
MTurk × Ability	0.059
MTurk × Age	0.003
MTurk × Gender	-0.060
CESS Online UK × Ability	-0.3
CESS Online UK × Gender	0.027
Ability × Age	-0.001
Ability × Gender	-0.032
Age × Gender	-0.002
Treatment × Online Lab × Ability	0.014
Treatment × Online Lab × Age	0.010
Treatment × Online Lab × Gender	-0.224
Treatment × MTurk × Age	-0.009
Treatment × MTurk × Gender	-0.052
Treatment × CESS Online UK × Ability	0.825
Treatment × CESS Online UK × Age	-0.013
Treatment × Ability × Age	-0.006
Treatment × Ability × Gender	0.096
Treatment × Age × Gender	0.007
Treatment × Age.2	0.0003
Intercept	0.563
ATE	-0.070

A CATE is estimated for each subject based on the model presented in Table A3 and their individual vector of treatment and covariate values. Figure A5 summarizes the estimation results. The horizontal blue line indicates an overall ATE of -0.07. 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 A5 presents the count of each mode for which the corresponding treatment effects in the upper part of Figure A5 is estimated. The mode histogram displays the distribution of subjects' mode along the spectrum of estimated treatment effect values. Almost all of the subjects who played the game in the lab are in the negative side of the CATE distribution, though their dispersion is wider than in the BART model. Subjects who played the game online are for the most part distributed to the right of those taking decisions in the lab, i.e. online CATEs have lower magnitudes. As in Figure 1, Mturk and CESS Online UK subjects' estimated treatment effects are predominantly located towards the right, positive end of the spectrum.

Figure A5: FindIt estimated heterogeneous effects including covariate interactions