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

May 6, 2019

[*] Funding and Acknowledgements: All of the experiments referenced in this manuscript were funded and conducted by the Nuffield College Centre for Experimental Social Sciences. All of the experiments received ethical approval from the Nuffield College CESS Ethics. Thomas Robinson is a recipient of an ESRC scholarship.

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 two diagnostic strategies, measurement metrics embedded in experiments and measurement experiments, that help assess whether any observed heterogeneity reflects experimental measurement error. To illustrate our argument, first 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. Secondly, we implement a measurement experiment in India with CESS Online subjects and 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). Short of outright fraud, experiments do not replicate either because of knife-edge treatment effects (Gelman, 2013) or experimental measurement error. 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). In this essay we demonstrate that experimental designs that incorporate diverse experimental modes can facilitate the detection of treatment effect heterogeneity and associated experimental measurement error.

Our novel contribution is to suggest how replicating identical experimental protocols across diverse experimental modes can inform efforts to detect experimental measurement error. We propose an iterative, machine-learning based estimation strategy in order to assess the magnitude of heterogeneous treatment effects across modes. These multi-mode micro-replications can be informative if researchers observe sufficiently different estimated treatment effects and they can distinguish, with reasonable precision, modes with high versus low experimental measurement error. We propose two further diagnostic strategies that help assess whether any observed heterogeneity reflects experimental measurement error. First, we propose to embed measurement items in the experimental protocol to help calibrate measurement error. Secondly, we advocate supplemental experiments that directly assess the magnitude of experimental measurement error. These diagnostic techniques along with multi-mode micro-replications help assess the robustness of estimated treatment effects.

We illustrate our case for multi-mode micro-replications with the results from 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 was 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] Separately, to illustrate how embedded, or complementary, experiments can help identify experimental measurement error, we conducted experimental vignette experiments in India with samples of 200 MTurk and 200 CESS India Online subjects.

We begin with a discussion of multi-mode micro-replications, suggesting why this strategy helps identify experimental measurement error. This is followed by three sections, each presenting a diagnostic strategy along with empirical examples: an iterative machine learning-based statistical method for estimating mode-specific heterogeneous treatment effects; measurement strategies for detecting experimental measurement error; and experimental approaches for evaluating conjectures about experimental measurement error.

2 Micro-replication and multi-modes

How should I micro-replicate? Many, if not most, experiments are conducted with a non-probability sample and can be implemented in a variety of modes. By modes we mean how the experimental treatments are delivered to subjects. Classic social science experiments are conducted in experimental labs where subjects receive treatments under the close supervision of the experimenter. Over the past decades, the modes for delivering experimental treatments have diversified dramatically. The subjects could be "workers" who agree to do paid tasks on the internet – MTurk being the most popular example although there are quite numerous variations on this theme. They could be part of a regular panel that agrees to answer various types of surveys on a regular basis. And experiments embedded in these surveys could be conducted in-person, online, on the phone, on various personal devices, or on Skype. Social media experimentation can take place on the internet with digital traces representing the outcomes of interest. And of course there is a proliferation of

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

experiments that are conducted in a wide-variety of field settings.

We contend that this diversity of experiment modes provides researchers with a unique opportunity to identify experimental measurement error that might undermine conclusions they draw regarding estimated treatment effects. Our contribution is to suggest what would be the most effective micro-replication strategy - given this diversity of experimental modes and constraints on researchers' resources. For many, having already invested in a particular mode – say MTurk – the preferred strategy is to replicate within mode.[2] We contend however that if the mode itself incorporates features that exaggerate, or contribute to, experimental measurement error then replicating within mode may not be particularly informative about fragile treatment effects. If the objective is to identify the measurement error associated with an experimental implementation, then the most cost effective micro-replication strategy is to invest in alternative modes.

The experimental endeavour is all about a design and implementation that will generate convincing results that replicate. Designs that incorporate micro-replications with multi-modes can help identify experimental measurement error that might call into question the results and make them difficult to replicate. The marginal payoffs of investing in alternative modes are higher than investing in the same mode. We suggest strategies that can be implemented with multi-mode micro-replications that help detect and diagnose experimental measurement error. As our simulations suggest, these efforts will be helpful under two reasonable assumptions: the prevalence of experimental measurement error varies by experimental context, or mode; and, secondly, when observing treatment effects that vary by mode researchers can typically identify the mode exhibiting lower experimental measurement error.

Experimental Measurement Error. Experimental measurement error occurs when subjects make choices or decisions that are an unintended artifact of the experimental design. This is a violation of the exclusion restriction since elements of the treatment delivery (mode)

[2] Note, that since our concern is with replication, we do not assume that researchers will be able to randomly assign mode (although this may be the experimental ideal).

are confounded with treatment effects. An underlying theme of much of the voluminous literature on experimental modes is the claim that experimental measurement error is, or is not, exaggerated in one mode versus another. Experimental measurement error fuels the debate in economics regarding the merits and failings of classic lab experiments as opposed to field experiments (Camerer, 2015; Levitt and List, 2015). One of the most widely-cited examples of experimental measurement error is the experimenter effect (Zizzo, 2010) that has been attributed to classic lab settings (Levitt and List, 2007), field experiments (Al-Ubaydli et al., 2017; Dupas and Miguel, 2017) and survey experiments (Bertrand and Mullainathan, 2001; Gooch and Vavreck, 2019), although de Quidt, Haushofer and Roth (2018) suggest its overall effect might be exaggerated. Online experiments have a range of potential experimental measurement errors that are unique to the mode, including inattention, online use of search browsers, trolling, and taking experiments multiple times. MTurk online experiments, particularly because of their popularity, have come under scrutiny due to potential measurement error associated with experimenter effects, the active social networks that link MTurk workers, questions about the nationality of MTurk workers, and, in fact, whether many MTurk workers are real people or simply bots (Kennedy et al., 2018; Burleigh, Kennedy and Clifford, 2018). Our general point is that regardless of what experimental mode is selected – a lab experiment, crowd-sourced worker experiment, online with highly paid subjects, lab in the field, Facebook-recruited subjects, online with representative panel, random control trials in the field – there will be well-regarded published authorities demonstrating the extent to which a particular mode is prone to experimental measurement error.

Researchers should adopt designs that maximize their chances of being informed about this potential experimental measurement error. By deliberately varying the experimental mode, researchers increase the likelihood of observing heterogeneous treatment effects that might be the product of measurement error. The challenge for researchers is discerning whether results are artifacts of experimental measurement error when they observe different mode-related treatment effects.[3]

Multi-modes Facilitate Measurement Error Detection. 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 implementation of multi-mode experiments can anticipate, and help account for, this measurement error (Loomes, 2005). Here, we revisit the estimator for treatment effects suggesting how multi-mode experimental design can facilitate the detection of experimental measurement error. The typical linear representation of the treatment effect for individual i on outcome y_i is:

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

where $\mathrm{Var}(\epsilon_i) = \sigma^2$ 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 $k \in K$ is the specific experimental mode, and T is the treatment variable.

The parameter values $\delta_k, \theta_k,$ and u_{ik} 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 – for instance, unclear instructions from the experimenter, inattentive subjects, credibility of payments of earnings. And, while this measurement error might generate considerable noise in the decisions made by subjects, it is random. Hence we only obtain imprecise estimates of the average treatment effect (ATE). 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 ATE will be biased towards the null. If some $\theta_k > 0$

[3] Recent findings by Bader et al. (2019) regarding the transportability of classic laboratory experimental findings to other modes highlight the issue of what constitutes mode-specific heterogeneity. In their case they clearly find that quantitative estimated treatment effects vary significantly across modes. On the other hand, the qualitative results are for the most part consistent across modes.

then we misreport measurement error for a treatment effect. This will result in an inflated estimate of the treatment effect.

Typically, we only observe y^*_{ik} for a single mode, k, and hence have limited information as to whether any of these three 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). A signal of experimental measurement error is heterogeneity in treatment effects across modes.

Our expectation is that if there is measurement error then it will likely vary by mode and hence δ_k and θ_k 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 significantly 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. Instead, 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.

Are multi-mode replications informative? Micro-replications are designed to identify experimental measurement error. If a micro-replication has a low probability of signaling measurement error it is probably not worth the investment. As we argue above, repeated sampling from the same mode may simply confirm the underlying bias created by mode-related measurement error. But of course, depending on one prior's regarding the prevalence of measurement error in different modes, multi-mode replications might not be cost-effective. We've constructed a simple simulation based on reasonable priors regarding the state of experimental measurement error. It suggests that multi-mode replication has as-good if not better payoffs than single-mode replication even when researchers' capacity to identify modes with low experimental measurement error are low.

We assume that there is some "true" average treatment effect in the world (ATE_T). This is the effect of some treatment, excluding any bias created by mode-related measurement error – attenuation biases, experimenter effects, falsified responses, and so on.[4] ATE_T is unobserved. A typical experimenter will observe ATE_k^*, where k is the mode in which the experiment is conducted. In our illustration, researchers can be in one of three "mode" states with some fixed probability, p_k such that $\sum_{k=1}^{3} p_k = 1$. ME_k is the experimental measurement error associated with mode $k \in 1, 2, 3$. In our illustration this can vary between 0 and 50. Each treatment effect will be observed with some probability: $ATE_k^* = (ATE_T + ME_k) \times p_k$. A key consideration in engaging in multi-mode micro-replication is the researcher's ability to detect which mode has a measurement error advantage. We capture this in the expected value calculation with the μ term. μ is the probability, conditional on observing the mode with the least experimental measurement error, of correctly identifying this mode, k, as the one with the least experimental measurement error.

We assume in this illustration that $\sum_{k \neq 2} ME_k > ME_2 = 0$. Researchers observe ATE_k^* and then make a decision regarding micro-replication; to either replicate within the same mode or to replicate with a different mode. Their replication decision determines their final estimated ATE_k^*. For those who choose the path of micro-replication within the same mode, their expected ATE_k^* will simply be a function of the p_k and ME_k. In a three mode setting, the expected measurement error for micro-replication within the same mode is as follows:

$$E(P, ME) = p_1 \times ME_1 + (1 - p_1 - p_2) \times ME_3 \tag{3}$$

Similarly we can compute the expected measurement error for those who opt for micro-replication in a different mode. Their expected observed treatment effect is slightly different. First, the expected value calculation will include the probability of ending up in one of the other two modes. It of course also includes the mode-specific ME_k. We include a third term, μ that is specific to those instances in which the researcher observes ATE_T, i.e, when $k = 2$ in our scenario. This term reflects the ability of the researcher to correctly identify $k = 2$ as

[4] Note since this is an average treatment effect, we are concerned with the mean response to treatment across the relevant population. The 'true' treatment effect for any subgroup may differ due to covariate factors that are not related to mode itself (which we explore in more detail in later sections).

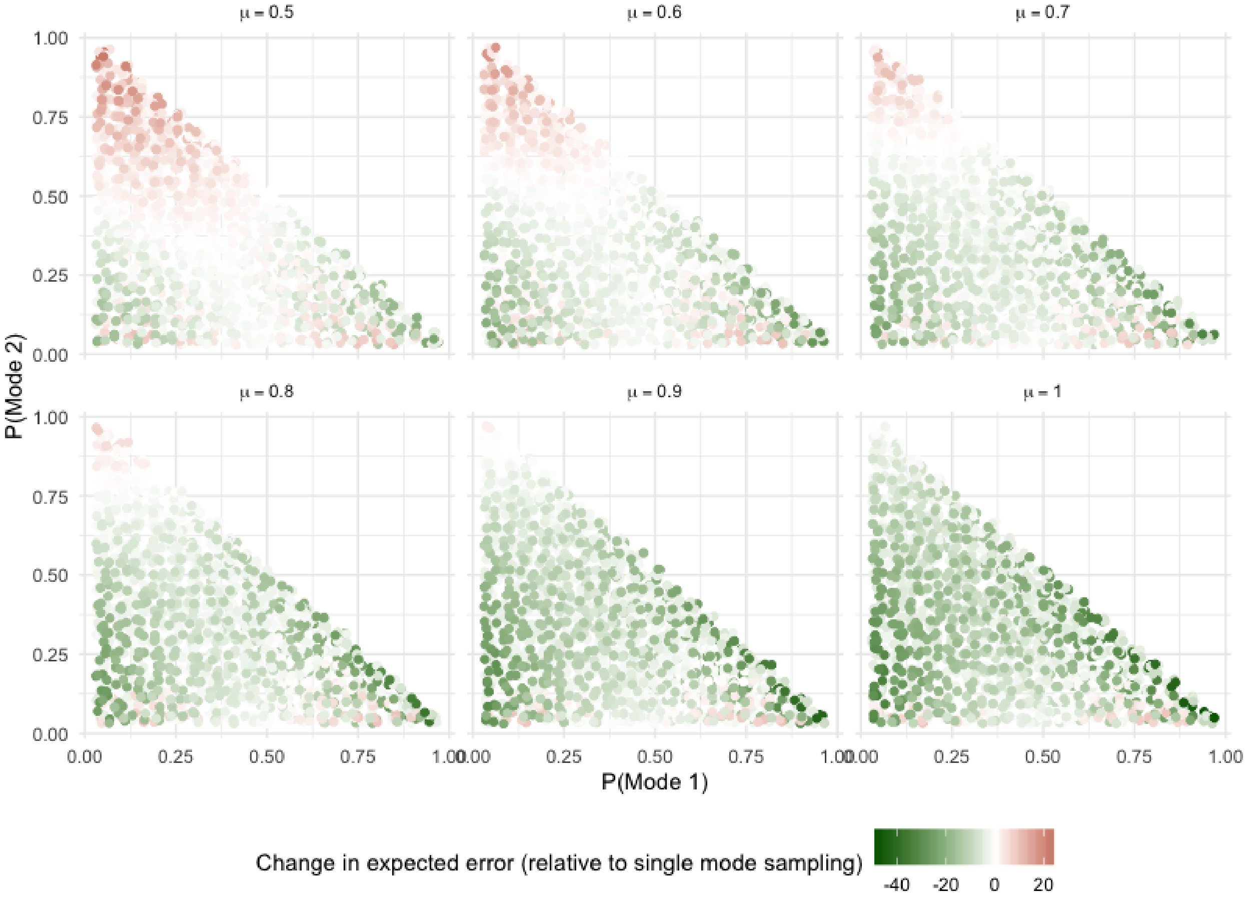
having $ME = 0$. Essentially this term reflects the success of efforts by researchers to embed measures and design micro-experiments that identify experimental measurement error. For the sake of simplicity in this illustration we assume mode comparisons are only informative if the researcher is comparing a mode with experimental measurement error to one without. In fact, the result can be generalized to cases in which researchers are comparing modes with varying levels of experimental measurement error. The expected measurement error for micro-replication in different modes is:

$$\begin{aligned}
E(P, ME, \mu) =& p_1 \times \frac{p_2}{1-p_1} \times ((1-\mu) \times ME_1) \\
&+ p_1 \times \frac{1-p_1-p_2}{1-p_1} \times \frac{ME_1 + ME_3}{2} \\
&+ p_2 \times \frac{p_1}{1-p_2} \times ((1-\mu) \times ME_1) \\
&+ p_2 \times \frac{1-p_1-p_2}{1-p_2} \times ((1-\mu) \times ME_3) \\
&+ (1-p_1-p_2) \times \frac{p_1}{p_1+p_2} \times \frac{ME_1 + ME_3}{2} \\
&+ (1-p_1-p_2) \times \frac{p_2}{p_1+p_2} \times ((1-\mu) \times ME_3)
\end{aligned} \tag{4}$$

The optimal micro-replication strategy depends on the three parameters: p_k, ME_k and μ, as Figure 1 illustrates. Here we have generated a schedule of outcomes for varying degrees of measurement error: $ME_k \in \{5, 10, 40\}$, for $k \in \{1, 3\}$; $ME_2 = 0$. And μ is varied in the interval $[0.5, 1]$; at $\mu = 0.5$, the researcher places equal weighting between the two modes i.e. has no relevant priors that lead to the researcher weighting their estimation in favour of Mode 2.[5] The values presented in Figure 1 are the effect on the expected measurement error of a multi-mode replication strategy relative to mono-mode replication. We vary the value of the measurement error in Modes 1 and 3 as detailed above, as well as the probabilities p_1

[5] It is possible to supply values of $\mu < 0.5$. These values, however, would suggest the researcher is biased against the zero measurement-error mode, which runs contrary to the assumptions of our basic model. For this reason, we focus only on those scenarios where the researcher is at worst ambivalent to Mode 2 ($\mu = 0.5$), or is biased towards Mode 2 ($\mu > 0.5$).

Figure 1: Simulation



and p_2 (and so, by implication, the probability of being in Mode 3: $p_3 = 1 - p_1 - p_2$).[6] Each facet represents a different value of μ, ranging from 0.5 to 1.0.

[6] Figure 1 shows the results for all the combinations of measurement error. The principle influence of measurement error in this model is to affect the magnitude of the payoff.

The red shading represents states of the replication world in which it is preferable to adopt a mono-mode replication strategy. First, and this is clear from Equation 4, as μ declines from certainty and approaches 0.5, the mono-mode replication strategy becomes optimal across a larger range of scenarios. This corresponds to research designs that are poorly equipped to detect experimental measurement error – hence unable to distinguish a treatment effect with minimal versus considerable experimental measurement error. Being unable to weigh more favorably the estimate within Mode 2 means that in expectation its often more advantageous to micro-replicate within the same mode. And we can see from Figure 1 that as a researcher becomes increasingly likely to recognize Mode 2 as having low measurement error, a high μ, then adopting multi-mode micro-replication is clearly the dominant strategy.

But for intermediate values of μ, which are probably the most plausible, i.e., between 0.5 and 0.7, we get some sense of how variations in the state of experimental measurement error affects micro-replication strategies. First, the probability of Mode 2 has to be much greater than 0.5 for it to make sense, in general, for researchers to adopt mono-replication strategies. Even when researchers have a greater than 50 percent chance of already having conducted an experiment with no measurement error, the fact that researchers will place greater weight on the zero measurement-error Mode 2 if it is one of the two modes chosen means that multi-mode replication becomes less costly in expectation.

Moreover, if the probability of a no experimental measurement error mode (i.e., Mode 2 with $ME = 0$) is less than 0.5 and μ ranges between 0.5 and 0.7 (i.e., the top row) then multi-mode replication is clearly the dominant strategy. Because the researcher can expect to observe one of two modes with measurement error *and* is more-than-likely to distinguish the no experimental measurement error mode (Mode 2) from the non-zero mode (either Mode 1 or 3), in expectation the multi-mode replication strategy will have lower experimental measurement error than a mono-mode replication.

The bottom row of 1 is in some sense aspirational. It represents a world in which researchers are almost certain to incorporate into their experimental design mechanisms for confidently detecting experimental measurement error ($\mu > 0.7$). In this case, a multi-mode micro-replication is virtually always the dominant strategy. The remainder of the essay suggests strategies both in the design, and subsequent analysis of results, that facilitate detection of experimental measurement error, i.e. increase the magnitude of μ.

3 Identifying Heterogeneous Mode-Effects

The benefits of multi-mode replications illustrated in Figure 1 are very much contingent on our ability to identify heterogeneous treatment effects that are associated with mode. And by mode here were are referring to features of the experimental design that determine how treatments are administered to subjects. In a multi-mode replication we treat experimental modes like covariates that could potentially condition treatment effects. There are potentially many other competing covariates that could be conditioning treatment effects. And, as Huff and Tingley (2015) suggest, clustering of particular covariates in different experimental modes could be confounded with what we are characterizing as mode effects.[7] The diagnostic utility of multi-mode replications depends on our ability to tease out the relative importance of mode in conditioning treatment effects.

To determine whether there are significant mode effects we implement an iterative, machine learning-based statistical method designed for estimating heterogeneous treatment effects (Athey and Imbens, 2017; Grimmer, Messing and Westwood, 2017; Künzel et al., 2019). 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 experimental modes. To the extent that there is no significant mode-related heterogeneity in conditional average treatment effects we gain some confidence that the estimated treatment effects are

[7] Although Coppock, Leeper and Mullinix (2018) compare survey experiment results from different modes and suggest this might not be an issue.

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-based statistical methods to estimate heterogeneity in treatment effects associated with these modes.

Experiment. We illustrate this approach using treatment effects generated from four different mode replications of identical experiments conducted by Duch, Laroze and Zakharov (2018). The overarching aim of this study was to understand how lying behaviour varies by economic context. These experiments consisted of lying games 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.

The lab experimental sessions were conducted at Nuffield CESS in Nov-Dec 2013 and Aug-Sep 2017. 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. Further details of these experiments are provided in the Appendix and in Duch, Laroze and Zakharov (2018).[8]

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.[9] In the lying module there is either a 0% or 10% audit rate that is fixed

[8] The complete replication material for Duch, Laroze and Zakharov (2018) is available at `https://github.com/rayduch/Once-a-Liar`. Replication material for the their specific lab experiment employed in our analysis is available at `https://github.com/rayduch/Experimental-Modes-and-Heterogeneity`.

[9] The decision to drop the second module is based on non-random attrition concerns – a substantive problem in experimental outcomes (Gerber and Green, 2008). While lab experiment subjects can reasonably be expected to stay in the lab for one or two hours (Morton and Williams, 2009), it is difficult to maintain

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).[10]

Treatment Effects. 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.

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.

participants attention online for that long (Mutz, 2011).

[10] All of the online material is available on the replication site: `https://github.com/rayduch/Experimental-Modes-and-Heterogeneity`.

	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

Heterogeneous Mode Effects Table 1 suggests a straightforward estimation strategy for identifying heterogeneous mode effects. We run separate GLM models for each mode, and there clearly is a pattern in the estimated coefficients suggesting 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 covariates are significant in some, although not all, modes.

The GLM estimation in Table 1 may be a perfectly reasonable specification for a model explaining lying behavior. It may not be the most conservative strategy for identifying heterogeneous treatment effects, however. The possible complication is that we are effectively imposing a particular specification ('ad hoc variable selection') that might not be optimal for identifying heterogeneous mode effects (Imai and Ratkovic, 2013, p.445). At least with respect to estimating possible heterogeneous mode effects typically we have no a priori expectations as to how mode interacts with either the treatment or other covariates. Nor do we necessarily have any priors on how other covariates interact with treatment itself.

To avoid imposing a restrictive structure on the estimation of the treatment and covariate 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 identical experiments using an iterative, machine learning-based statistical method. 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; Wager and Athey, 2018; Grimmer, Messing and Westwood, 2017). These techniques allow 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 value 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 machine learning-based strategy we believe is particularly suited to estimating mode-related heterogeneity: Bayesian Additive Regression Trees (BART) (Green and Kern, 2012; Hill, 2011).

BART is a frequently employed non-parametric modeling strategy that is simple to implement. It is the 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.[11]

To summarise, BART has several advantages over alternative estimation strategies. First, as with other supervised-learning methods, it takes the substantial decision about the functional form of the model out of the hands of the researcher. As the number of potential covariate-treatment interactions increases, the benefits of automating this aspect of estimation are greater (given the greater chance of model misspecification). Second, and in comparison to other tree-based supervised-learning methods, BART results are relatively robust to experimenters' choice of pruning parameters (Green and Kern, 2012). As a result, observed heterogeneity in CATEs across modes is less likely to be the result of idiosyncratic parameter selection. Third, even when outcomes are linear with treatment, BART's per-

[11] There are, of course, other strategies researchers can use to test for mode-related heterogeneous effects. Results for an additional estimation strategy, FindIt, are reported in the Online Appendix (Table A3 and Figure A5) – the findings are essentially the same as those reported for the BART method in this section. One could alternatively pursue exact matching on subject covariate values to estimate differences in outcome based on mode assignment. Of course, exact matching requires a sufficient number of subjects across modes to be effective. And, as a further complication, the researcher would have to match across both mode-assignment, and treatment assignment too. This strategy is beyond the purview of this essay, though we encourage others to pursue its viability. We return to the relevance of BART to other experimental contexts in the Discussion.

formance is very similar to the results of linear models (Hill, 2011). As we demonstrate below, even with a relatively small number of covariates, the results of the BART procedure nevertheless confirm the intuitions of the separate GLM models in Table 1. The greater flexibility of BART compared to GLM models, given its automated detection of treatment interactions, means this strategy can easily be scaled to larger sets of covariates. Finally, BART more easily enables us to recover individual CATE estimates, and to visualise mode heterogeneity in an informative way, which we turn to now.

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 $N \times C$ 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. The 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 counterfactual 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 corresponding observation in the synthetic, test dataset.

A CATE is estimated for each subject based on their individual vector of treatment and covariate values. This specification uses the same covariates as in Table 1 – age, gender, and ability rank – except here we pool observations across mode rather than estimating individual mode-specific models. Consequently, the ability rank covariate is now calculated with respect to the entire pooled sample, rather than individually within each mode. 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.[12]

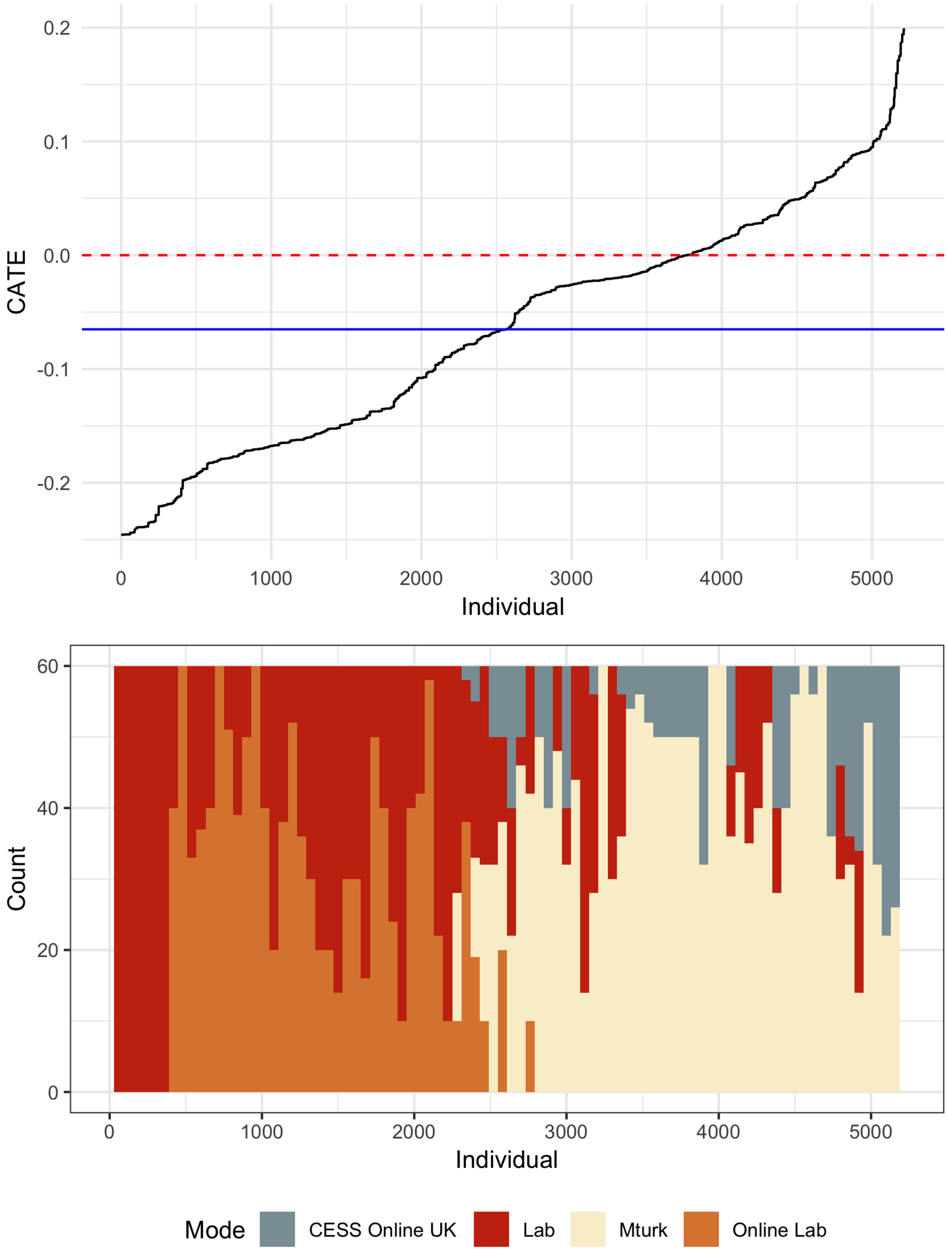
The distribution of CATEs organized by magnitude along with the histogram of covariate and mode/sample pool profiles are presented in Figure 2. The overall average ATE is -0.07 and the range over all the covariate values is -0.25 to 0.20. The distribution of CATEs generated by BART suggests that about 70 percent 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 -0.07 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 2 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 -0.07. Online subjects from either MTurk or CESS Online for the most part had CATEs greater than -0.07; and over half of these subjects (MTurk or CESS Online) 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

[12] The R code and all of the data employed here is available on the replication site: https://github.com/rayduch/Experimental-Modes-and-Heterogeneity.

Figure 2: BART estimated heterogeneous effects by mode



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 Online Appendix, result in very similar conclusions: *treatment effects differ by mode.*

Implementing multi-mode replications along with automated iterative estimation is a powerful diagnostic tool for identifying potential mode-related experimental measurement error. But since subjects in our, and most typical, experiments are not randomly assigned to modes (unlike Gooch and Vavreck, 2019), resulting evidence, or lack of, should only be treated as an initial, and indirect, indicator. The concern, of course, is that there are variables (unrelated to experimental measurement error), that we have not accounted for in the estimation, *and* that covary with mode. In our illustration for example, those subjects who participate via MTurk are unlikely to be similar to student participants who sign up for in-lab experiments. Our BART estimation strategy does include relevant covariates like age and gender that mitigate mode-related effects being confounded by demographic biases across mode. But of course these do not exhaust the subject characteristics that might be confounders here.

This particular diagnostic tool should be the point of departure – it establishes the likelihood of (or absence of) mode-related experimental measurement error. We suggest two subsequent diagnostic phases that assess whether any mode-related heterogeneity signals actual experimental measurement error.

4 Experimental Measurement Error

The first diagnostic phase of a multi-mode replication design is identifying mode-related heterogeneity using an iterative machine learning-based statistical method described above. Observing mode-related heterogeneity in CATEs is not particularly informative unless the research design incorporates explicit identification strategies. A second diagnostic component of the design, that we address now, determines whether mode-related heterogeneity actually signals experimental measurement error. We illustrate how embedded metrics in the experimental protocols can indicate whether mode-related heterogeneity reflects experimental measurement error – the illustrations focus on both random and systematic 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).[13]

The outcome variable in Duch, Laroze and Zakharov (2018), assessed in the previous section, 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).[14] An indicator of measurement error is 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

[13] 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 again 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.

[14] 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.

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.

Mode	(1)	(2)	(3)	(4)
Lab	0.769	0.905	0.76	0.85
	(0.028)	(0.02)	(0.047)	(0.032)
Lab Online	0.745	0.863	0.633	0.767
	(0.027)	(0.021)	(0.039)	(0.041)
CESS Onine	0.771	0.92	0.703	0.752
	(0.036)	(0.02)	(0.096)	(0.13)
MTurk	0.808	0.78	0.892	0.828
	(0.022)	(0.016)	(0.027)	(0.031)
Tax 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.

Mode	(1)	(2)	(3)	(4)
Lab	0.768	0.768	0.636	0.85
	(0.018)	(0.018)	(0.038)	(0.049)
Lab Online	0.807	0.76	0.762	0.767
	(0.017)	(0.017)	(0.02)	(0.047)
CESS Onine	0.88	0.827	0.827	0.752
	(0.011)	(0.017)	(0.026)	(0.029)
MTurk	0.758	0.758	0.782	0.828
	(0.016)	(0.012)	(0.024)	(0.027)
Tax Rate	10%	30%	10%	30%
Audited?	No	No	Yes	Yes

Table 3: Comparison of RET ICCs across modes

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 of subject performance across the four experimental modes. We employ the same strategy adopted 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	0.623	0.773
Lab Online	0.636	0.506	0.722
CESS Online	0.715	0.577	0.803
MTurk	0.749	0.688	0.794

Table 4: LTM Cronbach's alphas for integrity responses

There are mode effects as we demonstrated the previous section. But they do not seem to be associated with random measurement error. It does not appear to be the case, as some have suggested, that treatment effects are estimated much more imprecisely, or noisily, in the online mode, for example.

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 (e.g. Gooch and Vavreck, 2019). 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.

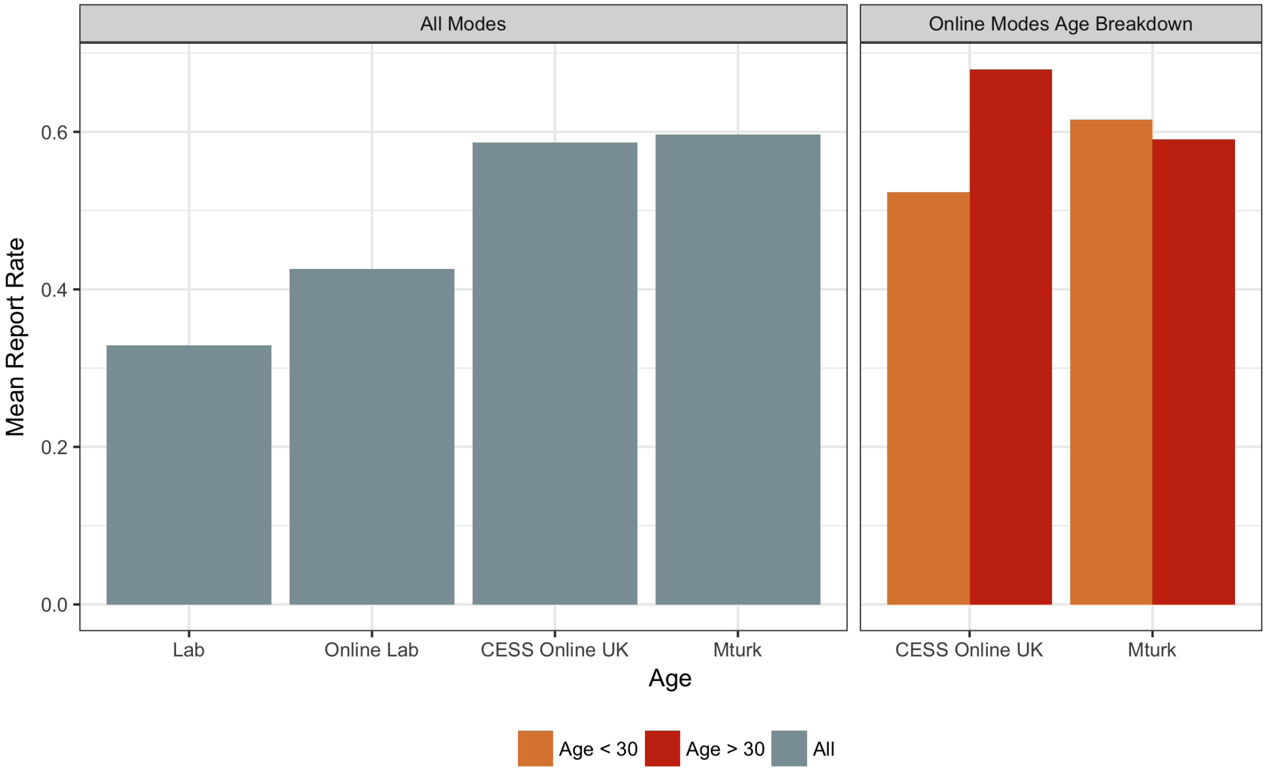
The challenge is incorporating features into the design that convincingly identify whether measurement error is generated by under-reporting (Blattman et al., 2016). Lying about earnings, the outcome variable in our experiment, is plausibly a sensitive choice for subjects to make. And there is an extensive 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 2 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 3 reports the incidence and magnitude of lying across the four modes.

There are two behavioral differences that stand out for the zero-audit condition in Figure 3. First, subjects, for the most part Oxford undergraduate, drawn from the lab subject pool (Lab and Online Lab) are more comfortable lying about their earnings. 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 2 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 3 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

Figure 3: Comparing Percentages of Actual Earnings Reported



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).

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 4 compares the distribution of age cohorts across different bands of CATEs for all four modes combined and also for just the MTurk mode separately. The estimated treatment effect is considerably lower in the case of the MTurk mode. The lowest treatment effect band for the MTurk mode straddles -0.05; in this band, there is no statistical difference in the proportion of those aged under, and over, 30 years old. In the case of the combined four

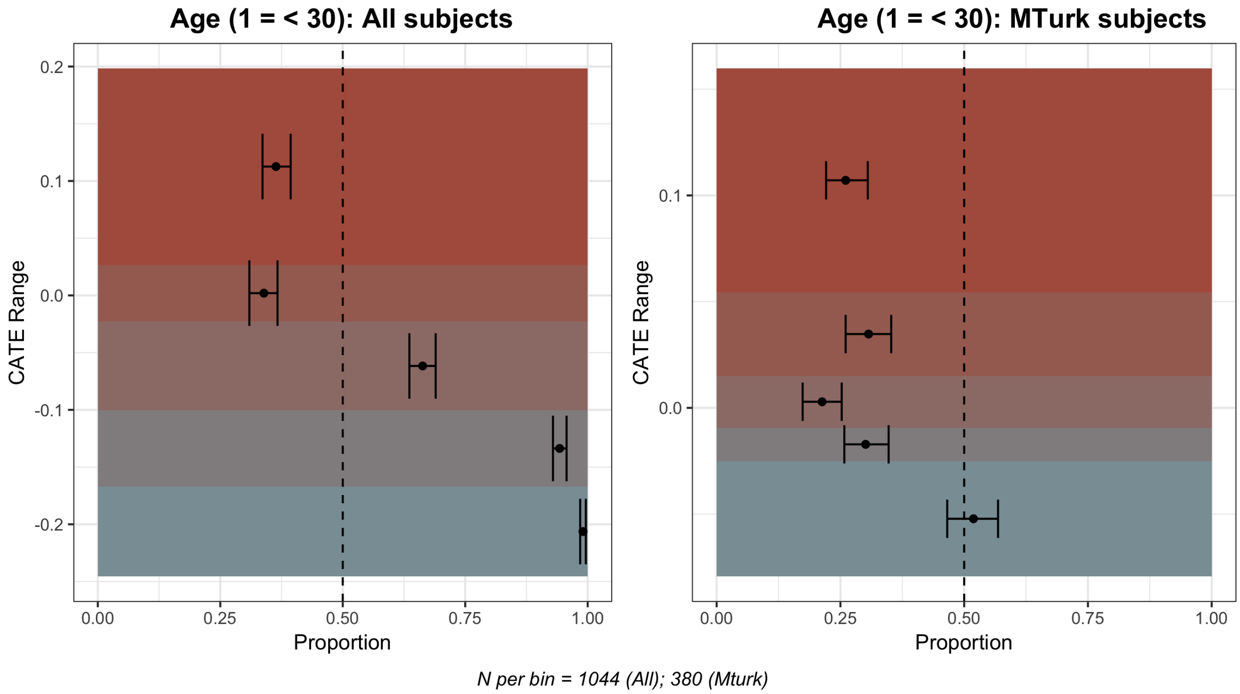


Figure 4: Banded bootstrap test for age covariate - all subjects versus MTurk only

modes we see treatment bands extending down to -0.20 which is substantively larger; and in this band we see a high concentration of young subjects. In both cases (all subjects and MTurk subjects alone), there is a larger proportion of older subjects observed for high CATE values (i.e. those approaching or exceeding 0) and a larger proportion of younger subjects for the lower CATEs. This suggests that the higher CATES estimated for the MTurk subjects is the result of a reluctance to lie on the part of both young and old alike.

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 2 are much larger for subjects from the Oxford lab subject pool, regardless of whether they played in the lab or online.

As part of the multi-mode replication design, the experimental protocol should anticipate this second component of the diagnostic process. Initial evidence of mode-specific heterogeneity simply signals the possibility of experimental measurement error. In this section, we proposed incorporating measurement metrics within the experimental protocol that allow researchers to assess the magnitude of measurement error in different modes.

5 Measurement Error by Design.

Our third micro-replication diagnostic explicitly incorporates measurement error into the experimental design in order to determine which experimental mode exhibits higher levels of measurement error. Again, ideally we want to be in the lower quadrant of Figure 1 where μ is high and we are almost certain to discern modes with high experimental measurement error. The notion of experimentally manipulating measurement error builds on similar efforts by de Quidt, Haushofer and Roth (2018), for example, who employ such manipulations to identify the bounds of experimenter demand effects on estimated treatment effects.

We propose a simple example to illustrate. Assume we observe treatment effects $ATE_{k=1}$ and $ATE_{k=2}$ in, respectively, Mode 1 and Mode 2. As we pointed out earlier, multi-modes might be informative if $ATE_1 \neq ATE_2$. You would like some evidence as to whether, and how, experimental measurement error is responsible for the difference in treatment effects.

Effectively there are four likely “states of the world”: 1) $ATE_1 = ATE_2$ which would be optimal (although even here both could be measured with error in which case the multi-mode design would be uninformative); 2) $ATE_1 < ATE_2$ or $ATE_2 < ATE_1$ and both ATE_1 and ATE_2 are measured without experimental measurement error (leaving unanswered the question why); 3) $ATE_1 < ATE_2$ where ATE_2 is measured without experimental measurement error although ATE_1 is not – in this case ATE_2 would be more plausible; 4) $ATE_1 < ATE_2$ where ATE_1 is measured without experimental measurement error although ATE_2 is not – in this case ATE_1 would be more plausible. Our focus has been on the latter two cases in which mode inequalities are potentially informative.

In our illustration, we observe $ATE_1 < ATE_2$ and our "μ-informed" priors are that ATE_2 has very little experimental measurement error. But of course we could be wrong – hence the need to incorporate design strategies that can substantiate the claim. By explicitly manipulating the measurement error associated with the experimental treatments we can gain insights into whether this is a plausible claim. Let $j \in \{\text{Neutral}, \text{High}\}$ be the extent of artificially manipulated measurement error. Thus, $ATE_{k,j}$ is the average treatment effect in mode k with artificially manipulated measurement error level j.

The diagnostic signals from these manipulations will be conditional on the prevalence, and nature, of experimental measurement error. A consideration is that experimental measurement error could either inflate or depress the estimated treatment effects.[15] In our illustration, we assume the latter; experimental measurement error is presumed to depress the estimated treatment effects. Mode 1 represents an experimental context in which treatments are administered with high levels of error. Effectively this means that fewer subjects are being "treated" and in our illustration we expect this will bias the estimated treatment effect to zero. Explicitly adding measurement error to treatments that already exhibit high measurement error will have relatively small effects on the estimated treatment effect. And any observed effect will be towards the null.

We could, and will in the example below, observe an $ATE_1 = 0$. On its own, this observation does not preclude the counterfactual that we are just observing a null treatment effect. Adding measurement noise to the treatment in this case should also have little effect. The more informative diagnostic in this case, then, is to include an additional treatment for Mode 1 that reduces the measurement error. If experimental measurement error in Mode 1 is suppressing the treatment effect then our expectation here is that a treatment reducing error would result in a more positive treatment effect.

Our priors regarding Mode 2 are that it has low levels of experimental measurement error. The complementary diagnostic that will be informative here is a treatment that also

[15] The classic "experimenter effect" for example could have an inflationary impact on estimated treatment effects (de Quidt, Haushofer and Roth, 2018).

increases measurement error. Adding a significant amount of measurement noise to Mode 2 is expected to reduce the magnitude of the treatment effect. We should observe a substantial reduction in the treatment effect when we add noise to a treatment context that *a priori* has little measurement error.

As an illustration we implemented such a design strategy as part of the Nuffield Centre for Experimental Social Sciences (CESS) 2019 Vote India election study. The simple survey vignette experiment aims to assess the ability of the Indian general public to identify fake election news stories. The theme of the news stories was the reliability of electronic voting machines used to tally votes in 2019 Indian Lok Sabha elections.[16] We implemented the experiment in two different modes during the period April 2-26, 2019: MTurk workers from India (our Mode 1), and the CESS India Online subject pool (equivalent to our Mode 2). As part of the initial design we included treatment assignments to identify heterogeneous mode effects that could be associated with measurement error. Again, our primary point here is to leverage multi-mode micro-replications in order to learn about experimental measurement error.

We implemented a "neutral" version of the vignette experiment with a control consisting of an accurate news statement randomly assigned to approximately 50 MTurk and 50 CESS Online subjects, respectively. A "neutral" version of the fake news treatment was also assigned to approximately 50 MTurk and CESS Online subjects. We also implement a "high-error" version of both the accurate news statement and the fake news treatment. Again, each of these versions was randomly assigned to approximately 50 MTurk and 50 CESS Online subjects. The vignettes for each group are displayed in Table 5. What varies here is the deliberate framing of the treatment. The "neutral" version is designed to minimize the likelihood of any confounding measurement error. In the "high-error" version we deliberately incorporate framing elements that we expect will create measurement error in fake news detection.

[16] Voting machine reliability was a frequent theme in news accounts of the Lok Sabha elections. https://www.bbc.com/news/world-asia-india-46987319.

	Control	**Treatment**
Neutral	The Indian Election Commission has announced that the coming Indian elections will continue to use electronic voting machines.	The Indian Election Commission has warned that there is likely to be extensive election fraud in the upcoming Lok Sabha elections because of the use of electronic voting machines, that can be easily hacked.
High Error	The Indian Election Commission has said all polling booths will have the voter-verified paper audit trail facility this election, a system in which voters can see on paper whether the machine has registered the same vote as the button they pressed. This provides an additional layer of security and reduces the possibility of massive electoral fraud.	The Indian Election Commission has been ordered to discontinue the use of voter-verified paper audit trail facility this election, an old system in which voters could see on paper whether the machine had registered the same vote as the button they had pressed. Opposition parties have alleged that the BJP is behind this change and that not using paper trail will lead to massive electoral fraud.

Table 5: Vote India Vignettes by Treatment Group and Error Version

We implemented a third "attention-incentivized" version of the vignette experiment for the MTurk subjects. Subjects were asked to complete exactly the same fake news detection task as in the "neutral" and "high-error" conditions. But in this version respondents also saw the following text: "On the following page, after you indicate how truthful or false the statement is, we will then ask you a factual question about the statement itself. If you answer this factual question correctly, you will be paid an additional 25 INR." The factual question asked participants to select which institution was mentioned in the vignette text (i.e., the Indian Electoral Commission). Our goal here is to reduce measurement error resulting from inattention. We believed we could accomplish this by both signaling that there would be a treatment check (a factual question about the treatment) *and* incentivizing correct answers. The incentivised version of the experiment was fielded on 200 subjects. Of those respondents, 131 correctly identified that the question mentioned the Indian Electoral Commission. Even with financial incentives, over a third of the sample failed to pay sufficient attention to answer a straightforward descriptive question.

The conditions in Equations 5-6 summarize the information we expect to garner from this diagnostic exercise. Our priors here are that the CESS India Online subject pool would exhibit low levels of experimental measurement error. Accordingly, in the neutral condition detailed in Table 5 we expect to see a significant treatment effect. In the high error condition, this treatment effect will be smaller but significantly different than zero. For modes with little experimental measurement error the high-error condition should noticeably attenuate the treatment effect.[17]

$$ATE_{\text{MTurk,Neutral}} \geq ATE_{\text{MTurk,High}} = 0, \tag{5}$$

$$ATE_{\text{CESS,Neutral}} > ATE_{\text{CESS,High}} > 0. \tag{6}$$

The expected outcomes for the MTurk subjects are quite different. We anticipate that the MTurk mode will have considerable measurement error and accordingly in the neutral condition we expect to see a weak or possible null finding for the fake news treatment. And our expectations for the high error condition are not that much different: we expect a null result similar to the neutral condition. Our reasoning here is that for modes that already exhibit high experimental measurement error this exaggerated measurement error will have little effect on the estimated treatment effect (compared to the low-error condition).

For the MTurk subjects we implement a condition, for both neutral and high error treatments, in which subjects are incentivized to focus their attention on the fake news treatments – our "attention" condition. Let ATE'_{kj} be the average treatment effect for those participants in the attention condition, given mode k and with manipulated measurement error j as before. The expectation is that with additional incentives, the neutral and high error treatments for MTurks will be significant – hence more closely resembling the CESS India Online treatment effects. The hypothetical expectations are given in Equations 7 and 8.

[17] $ATE_{k,j} = E[Y_i | k_i = k, j_i = j, T_i = 1] - E[Y_i | k_i = k, j_i = j, T_i = 0]$.

Coefficient	S.E.	t-statistic	p	Mode	Error Type	Incentivised?
-0.74	0.47	-1.57	0.12	MTurk	Neutral	No
-0.83	0.47	-1.76	0.08	MTurk	High	No
-3.85	0.51	-7.52	0.00	CESS Online	Neutral	No
-3.23	0.49	-6.64	0.00	CESS Online	High	No
-1.16	0.49	-2.35	0.02	MTurk	Neutral	Yes
-1.00	0.33	-3.01	0.00	MTurk	High	Yes

Table 6: Induced measurement error – model results by mode, error type and attention condition

$$ATE'_{\text{MTurk,Neutral}} > ATE'_{\text{MTurk,High}} > 0, \tag{7}$$

$$ATE'_{\text{MTurk,Neutral}} > ATE_{\text{MTurk,Neutral}}. \tag{8}$$

The results for the non-incentivized conditions are reported in the first four rows of Table 6. There is a much greater treatment effect for CESS Online compared to MTurks subjects (MTurk < 1, and CESS Online > 3). A design relying exclusively on the MTurk mode would favor the null – the inability of subjects to detect fake news. But this is also consistent with our initial priors – that the MTurk treatments would have considerable measurement error depressing estimated treatment effects. On the other hand, and again consistent with our priors, the CESS Online treatment effects in rows 3 and 4 are large and statistically significant. Moreover, the treatment coefficients for all four models are essentially the same when we include controls for age and gender (see Appendix for full results). A possible conclusion here is that MTurk subjects are considerably less attentive than CESS Online subjects and hence were effectively not being "treated". The result is an insignificant treatment effect.

This intuition is reflected in the expected neutral/high error treatment effects presented in Equations 5 and 6. As expected, manipulating the measurement error in treatments had no affect on the MTurk treatment effects – the neutral and high error results are similarly small and insignificant. Again, as expected, we do see treatment effects drop (over one standard error) when we introduce identical measurement error for the CESS India Online subjects. These differential responses to the measurement error treatments suggest that the MTurk mode results are depressed by experimental measurement error. This could very well be the result of MTurker's inattention, or even MTurk bots that increase measurement error. We interpret the impact of the high error version of the treatment effect for CESS Online subjects as an indication of relatively subdued experimental measurement error – particularly in contrast to the MTurk results. Adding measurement to an experimental context with little prior measurement error should moderate the treatment effects although note they are still substantial and significant in the high error condition.

These contrasting results for MTurk versus CESS India Online strongly suggest that experimental measurement error is depressing treatment effects for the MTurk subjects. Our conjecture here is that it is measurement error depressing the MTurk treatment and our hunch is that at least some of the error is induced by inattention to the treatments. We assessed this conjecture by incentivizing attention to the treatments – Equations 7 and 8 presented our expected outcomes. And the final two rows of Table 6 display the results. The estimated ATE for both "neutral" and "high error" versions of the treatment are substantively larger than their unincentivised counterparts. Both coefficients are statistically significant, and the ATE of the "neutral" treatment is now moderately larger than the corresponding "high error" version. There is strong evidence here that by experimentally reducing MTurk inattention to treatments we obtain results comparable to those for the CESS India Online subjects.[18]

These fake news detection results illustrate our broader theme of the importance of designs that identify mode-specific heterogeneous treatment effects. Certainly in this case

[18] In fact, of the 131 individuals who correctly identified the Indian Election Commission, only 106 selected this option alone. The percentage of participants who failed to isolate this institution is still around 45 percent even with monetary incentives. Running the estimations on just those 106 participants, the high and low error ATE estimates both increase in size, remain statistically significant, and the difference in ATEs increases between the error levels too: Low error ATE = -2.16 (s.e. = 0.78); High error ATE = -1.70 (s.e. = 0.57).

it could be problematic to rely exclusively on the MTurk mode. Our specific goal in this section though is to suggest a third diagnostic strategy for determining whether mode-specific heterogeneous treatment effects are a product of experimental measurement error. Having observed mode-specific heterogeneity, we recommend designing measurement error experiments that directly assess the underlying source of the error. In our case, we designed treatments that explicitly manipulated the contextual features of the experiment that are claimed to cause experimental measurement error.

6 Discussion

Technology, ingenuity and cost have all contributed to the diversity and accessibility of experimental modes available to the average researcher. We should exploit this rich diversity of experimental modes in order to understand and address experimental measurement error in our replications. Most recognize that either your reported effect sizes or null effects will be an artifact of some feature of how, and in what context, treatment is assigned. The interesting challenge is to understand the source and magnitude of this experimental measurement error (de Quidt, Haushofer and Roth, 2018; Gillen, Snowberg and Yariv, Forthcoming). Our contribution in this respect is twofold: first we explain why multi-mode designs are informative about experimental measurement error and secondly provide suggestions for deploying multi-mode designs as a diagnostic tool.

We presume a best practice of incorporating micro-replication within the experimental design. A researcher is faced with the choice of a mono- versus multi-mode replication strategy. We assume there is some mode-related heterogeneity in experimental measurement error and researchers are reasonably adept at detecting this measurement error. With these quite reasonable assumptions, we demonstrate that multi-mode replication designs are clearly the most informative about experimental measurement error.

We suggest how experimental modes can be deployed as a diagnostic tool for detecting experimental measurement error. Our recommendation is to incorporate numerous diverse experimental mode replications in the design – in our example we had four distinct modes. One of the diagnostic contributions of the essay is a simple machine learning-based strategy that identifies heterogeneous treatment effects. The researcher imposes no *a priori* specification to the nature of the potential heterogeneity. CATEs are generated for all subjects across all experimental modes. We then show how mode-specific heterogeneity can be estimated and viewed graphically. The absence of mode-specific heterogeneity speaks to the robustness of the estimated treatment effects.

We advocate for estimating mode-related heterogeneity using BART as it is a flexible estimation strategy that can cater for a wide variety of experimental contexts. As the number of potential covariate-treatment interactions increases, the benefits of BART over linear regression strategies also increase. Given the ease of implementing BART estimation, its robustness to varying numbers of covariates, and its accurate predictions regardless of covariate length (Hill 2011), we believe the procedure should be used as the standard means of estimating both CATEs and mode-related heterogeneity.

That said, data-generating processes in specific experimental settings may mean BART is not the most appropriate tool. In these cases, we believe researchers should still pursue automated machine-learning strategies to avoid the problems associated with model misspecification. In the appendix, we demonstrate the results of one such alternative strategy: LASSO models using the Support Vector Classifier Imai and Ratkovic (2013). Other machine-learning estimation strategies include the use of ensemble methods (Grimmer, Messing and Westwood, 2017) and meta-learning strategies to handle imbalances in the number of subjects assigned to treatment and control (Künzel et al., 2019). In general, researchers should recognize that the appropriateness of machine-learning estimators is contingent upon the nature of their data and the data generating process, and choose their estimation strategy accordingly.

Using these automated strategies, observing different treatment effects across modes is a strong signal of mode-specific measurement error but hardly definitive. Multi-mode replication designs are particularly powerful if researchers are able to demonstrate that in fact mode-specific heterogeneity is (or is not) related to experimental measurement error. This is important because it allows the researcher to discriminate between possibly quite different treatment effects.

Accordingly, we suggest diagnostic tools that detect experimental measurement error. They are aimed at helping resolve the quandary of contradictory treatment effects observed for identical experiments administered in different modes. The experimental context, which covers a variety of factors relating to administering treatments and recording outcomes, affects, in either a systematic or random fashion, measures of both outcome and control variables. Embedding measurement items in experimental protocols can help determine whether a particular mode exhibits systematic or random measurement error. Our examples included measurement scales (Are respondents scaling as we would expect them to?); repeated measures (Are their answers correlated over time?); and indicators of sensitive questions (Are subjects underreporting certain behaviors or preferences?). One possible extension of this strategy would be to include metrics to detect measurement error in the experimental protocol that can be estimated using automated machine-learning strategies like BART.

Researchers typically have good intuitions regarding experimental measurement error. These should be deployed for designing measurement error experiments that might explain mode-specific heterogeneous treatment effects. A third diagnostic strategy builds on these intuitions, implementing experiments with treatments that manipulate levels, and types, of measurement error. The treatments, applied to the different modes, are designed to explore conjectures regarding experimental measurement error. In our fake news detection example, our conjecture was that inattention to treatments by MTurk subjects might contribute to experimental measurement error and hence depress the treatment effects. A series of measurement experiments that both exaggerated and minimized measurement error provided suggestive evidence to this effect. Our point here is that similar experiments informed by solid intuitions about the causes of experimental measurement error can help draw sensible conclusions when confronted with mode-specific heterogeneous treatment effects.

To conclude, it is worth reflecting on future experimental design and data generation. We have hopefully convinced readers that multi-mode micro-replications add value to experimental designs. Our contribution relates to the identification of adaptive experimental designs that integrate multi-mode data generation. One unanswered question is whether in the design phase there are optimal strategies for ensuring that researchers generate experimental data that maximizes their informative value regarding experimental measurement error. Researchers have increasingly diverse mode options for conducting experiments yet under-developed strategies for selecting amongst them, and for deciding how they might be combined in a complementary fashion. Moreover, one could envisage designs that also incorporated experimental treatments that were explicitly directed at helping measure, and possibly correct for, experimental error. A second, and related, challenge therefore is the detection and measurement of experimental measurement error itself. We propose two distinct estimation strategies for these two tasks. Again, in an effort to integrate this estimation, future work should explore machine-learning methods that combine both the estimation of heterogeneous treatment effects and any associated experimental measurement error.

References

Al-Ubaydli, Omar, John A. List, Danielle LoRe and Dana Suskind. 2017. "Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature." *Journal of Economic Perspectives* 31(4):125–44.

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

Bader, Felix, Bastian Baumeister, Roger Berger and Marc Keuschnigg. 2019. "On the Transportability of Laboratory Results." *Sociological Methods & Research* 0(0):0049124119826151.

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.

Bertrand, Marianne and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *Economics and Social Behavior* 91(2).

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.

Burleigh, Tyler, Ryan Kennedy and Scott Clifford. 2018. "How to Screen Out VPS and International Respondents Using Qualtrics: A Protocol." Working Paper.

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.

Coppock, Alexander, Thomas J. Leeper and Kevin J. Mullinix. 2018. "Generalizability of heterogeneous treatment effect estimates across samples." *Proceedings of the National Academy of Sciences* .

de Quidt, Jonathan, Johannes Haushofer and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* 108(11):3266–3302.

Duch, Raymond, Denise Laroze and Alexei Zakharov. 2018. "Once a Liar Always a Liar?" Nuffield Centre for Experimental Social Sciences Working Paper.

Dupas, Pascaline and Edward Miguel. 2017. Impacts and Determinants of Health Levels in Low-Income Countries. In *Handbook of Economic Field Experiments*, ed. Esther Duflo and Abhijit Banerjee. Elsevier pp. 3–93.

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.

Gerber, Alan S. and Donald P. Green. 2008. *The Oxford Handbook of Political Methodology.* Oxford University Press chapter Field Experiments and Natural Experiments, pp. 357–381.

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

Gooch, Andrew and Lynn Vavreck. 2019. "How Face-to-Face Interviews and Cognitive Skill Affect Item Non-Response: A Randomized Experiment Assigning Mode of Interview." *Political Science Research and Methods* 7(1):143–162.

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. “Estimating Treatment Effect Heterogeneity in Randomized Programme Evaluation.” *The Annals of Applied Statistics* 7(1):443–470.

Kennedy, Ryan, Scott Clifford, Tyler Burleigh, Ryan Jewell and Philip Waggoner. 2018. “The Shape of and Solutions to the MTurk Quality Crisis.” Working Paper.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel and Bin Yu. 2019. “Metalearners for estimating heterogeneous treatment effects using machine learning.” *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel and Bin Yu. 2019. “Metalearners for estimating heterogeneous treatment effects using machine learning.” *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Levitt, Stephen and John List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World.” *The Journal of Economic Perspectives* 21(7):153–174.

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.

Lupton, Danielle L. 2018. “The External Validity of College Student Subject Pools in Experimental Research: A Cross-Sample Comparison of Treatment Effect Heterogeneity.” *Political Analysis* pp. 1–8.

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.

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.

Mutz, Diana C. 2011. *Population-Based Survey Experiments.* Princeton 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.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.

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

Zizzo, Daniel John. 2010. "Experimenter demand effects in economic experiments." *Experimental Economics* 13(1):75–98.

Appendix A Lying experiment

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, the 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. Audit rates of zero and Tax rates of 10% can be considered the baseline categories for comparisons. Subjects in all modes completed a Dictator Game and a risk aversion lottery.

Mode	DG	Risk	Audit	Tax	Report	#	#	%	%
			Rate	Rate	Rate	Subjects	Obs	Female	Male
CESS Online UK	Yes	Yes	c(0, 0.1)	c(10, 30)	0.63	90	696	0.52	0.48
Lab	Yes	Yes	c(0, 0.2)	c(10, 20, 30)	0.43	116	1600	0.44	0.56
Mturk	Yes	Yes	c(0, 0.1)	c(10, 30)	0.60	390	2419	0.49	0.51
Online Lab	Yes	Yes	c(0, 0.1)	c(10, 30)	0.46	144	1367	0.50	0.50
All	Yes	Yes	c(0, 0.1, 0.2)	c(10, 20, 30)	0.53	740	6082	0.48	0.52

Table A1: Summary of experimental treatments

We find some differences across subject pools with respect to other-regarding preferences. In the classic Dictator Game (with a 1000 ECUs pie) 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 available in the Online Appendix).

We elicited risk preferences 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.

In the lying game subjects had to invest effort to earn money and make decisions about lying; they made these decisions in groups of four, in real time; and the groups 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 (Figure A4 in the Online 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.

A2 Sample Covariates

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 where there is a higher proportion of male subjects. As indicated in Figure A1 there are substantive age differences in the three subject pools. The student subject pool, used in the lab and online experiments, are younger than subjects from the online 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

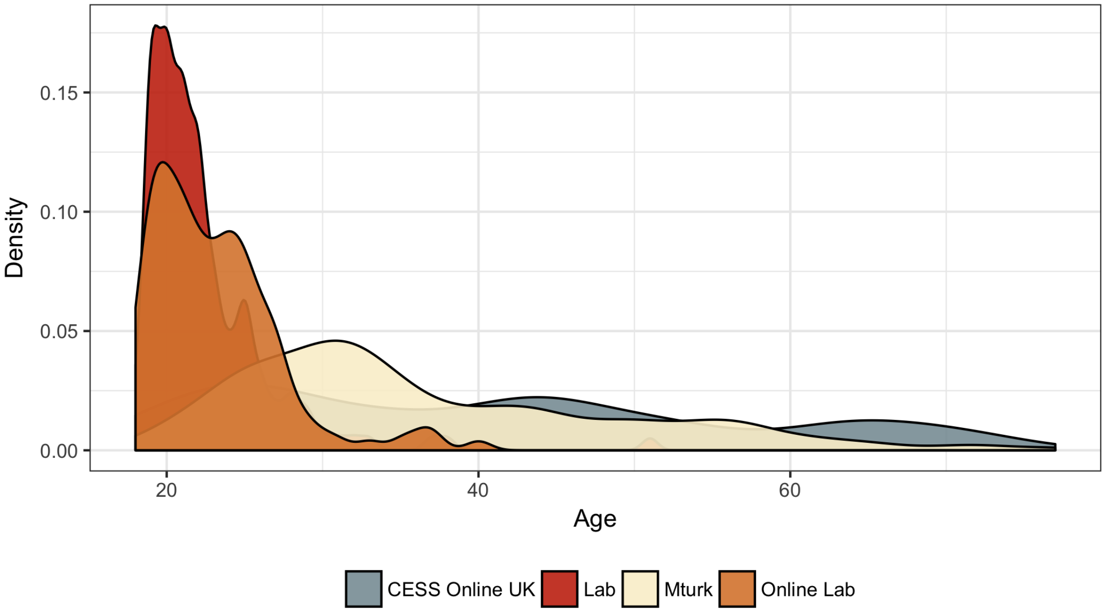


Figure A1: Age distribution of subjects

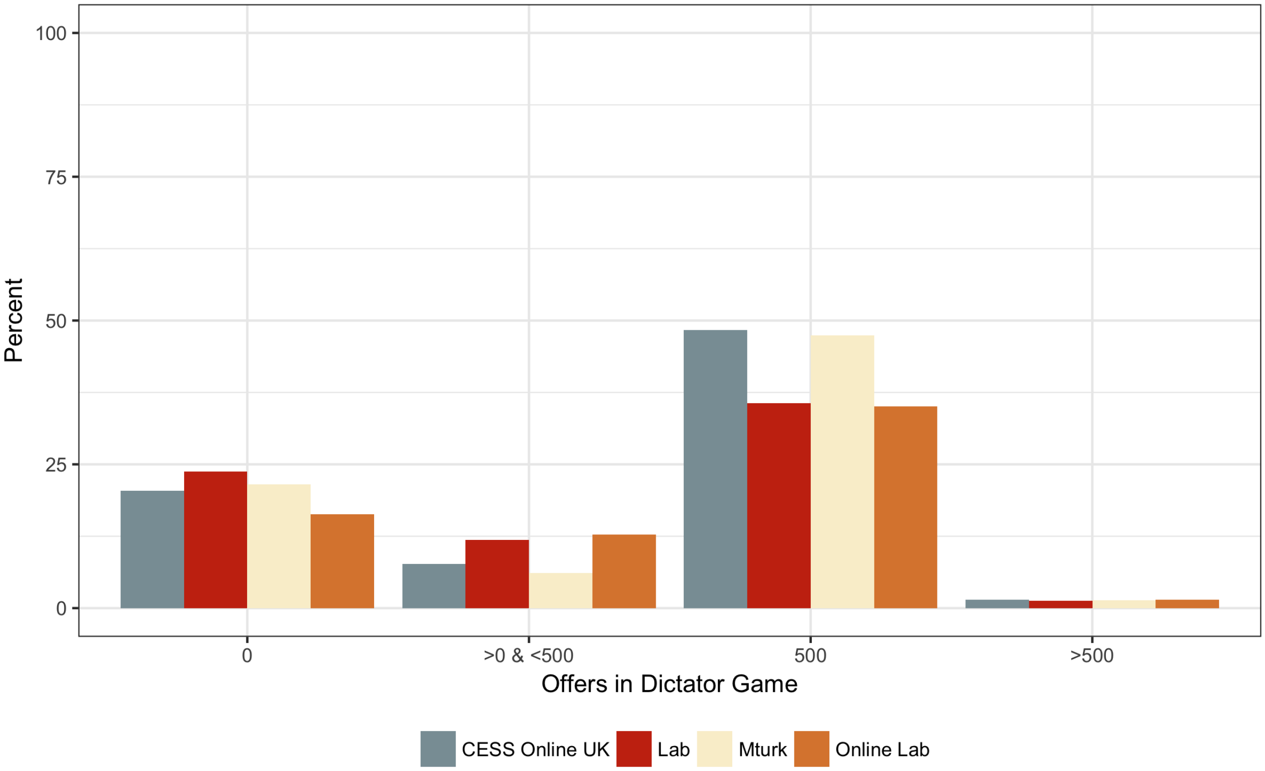
younger on average. The general UK online panel subjects are similar to MTurk subjects. The age distributions for MTurk and UK online are significantly different from UK lab and online in both t-test and Wilcoxon rank sum test, but MTurk and UK online are not distinguishable at the 95% confidence level (results in Online Appendix Table A1).

A3 Decision-theoretic preferences

One concern is that subject pools may differ with respect to fundamental preferences (Belot, Duch and Miller, 2015; Lupton, 2018). 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

Figure A2: Dictator Game



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 online, 329 by the general UK panel and 307 by Mturk workers.

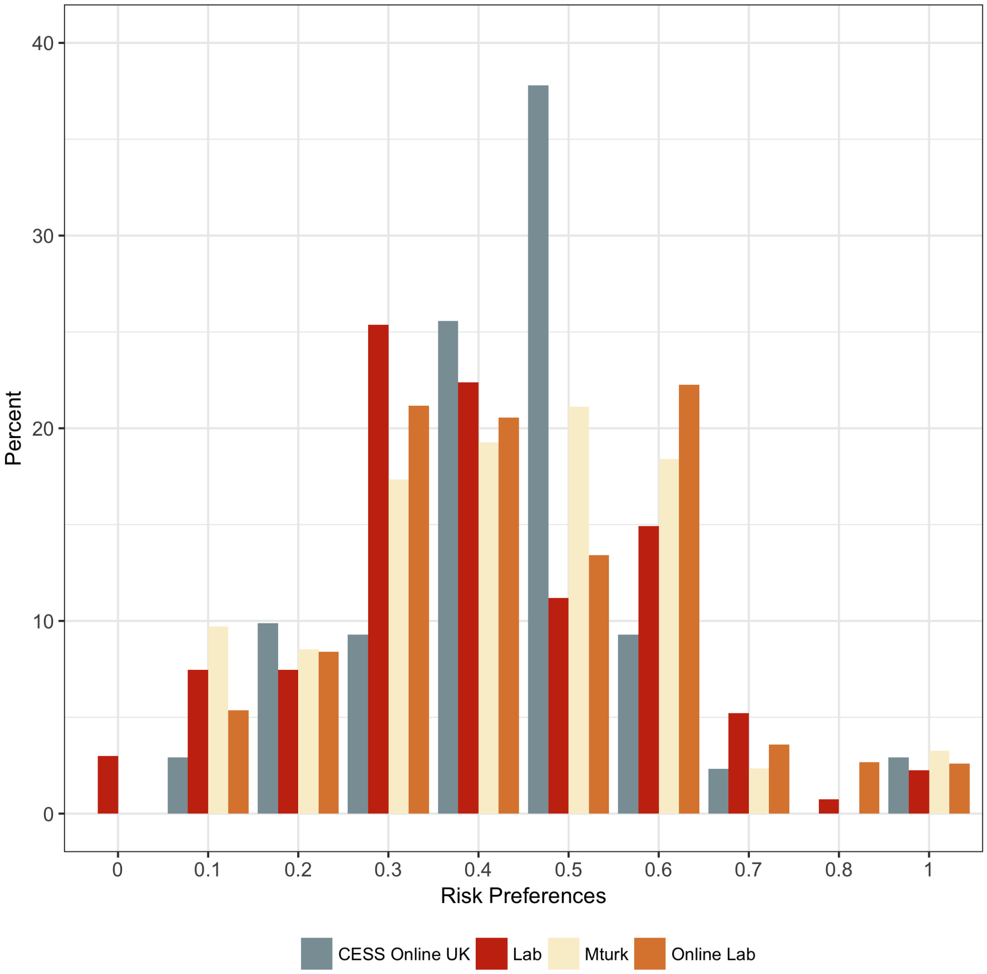
Students are more likely to offer nothing when they are in the lab, but in both *t*-test and Wilcox rank sum tests, the difference between students in the lab and online is insignificant. In contrast, the UK Online panel and Mturk subjects are significantly more generous than the two student subject 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 preferences, 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

Figure A3: Risk Preference



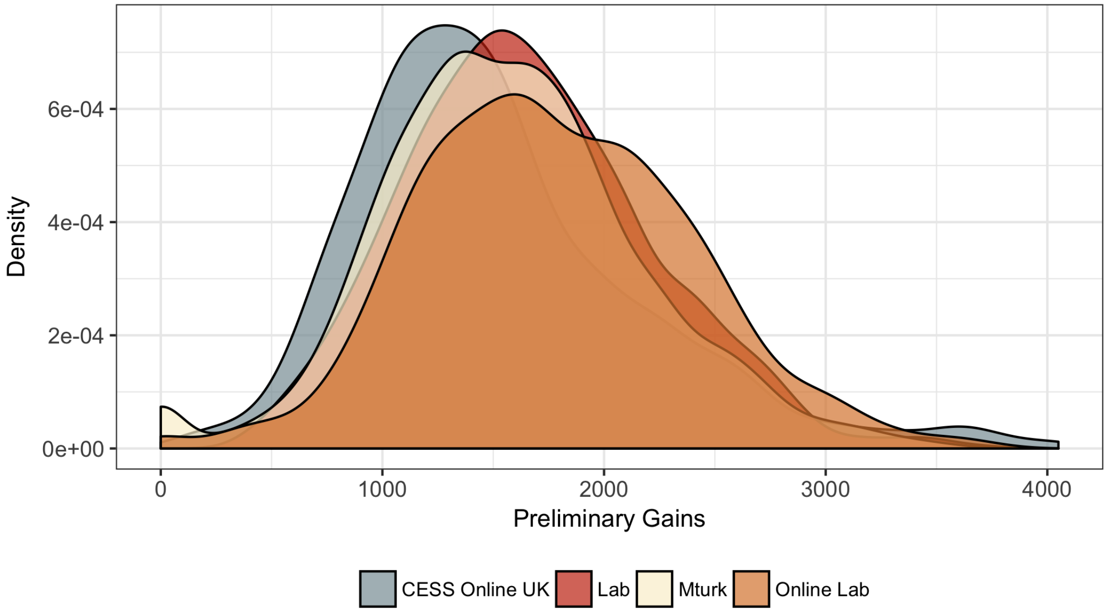
levels of significance ($p > 0.1$). And at least two of the three different online subject pools made very similar average offers in the Dictator Game. On the second incentivized risk preference experiment subjects made similar choices – none of the risks results from the four experiments suggested a significant mode or sample difference.

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			
		Online	Online			Online	Online	
	Lab	Lab	UK	MTurk	Lab	Lab	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). 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 – 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).[19] 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 and interactive effects is generated using the converged values of the LASSO tuning parameters.[20]

In our case, the LASSO model generated non-zero heterogeneous parameter estimates for subjects across mode conditions. This is particularly noteworthy given the sparse estimation strategy of LASSO models. From this model, we then predict the expected effect of treatment for each individual's sample, mode and treatment assignment plus their vector of covariate values using the inbuilt predict function within the FindIt package (see Egami, Ratkovic and Imai, 2018).

[19] 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.

[20] 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.

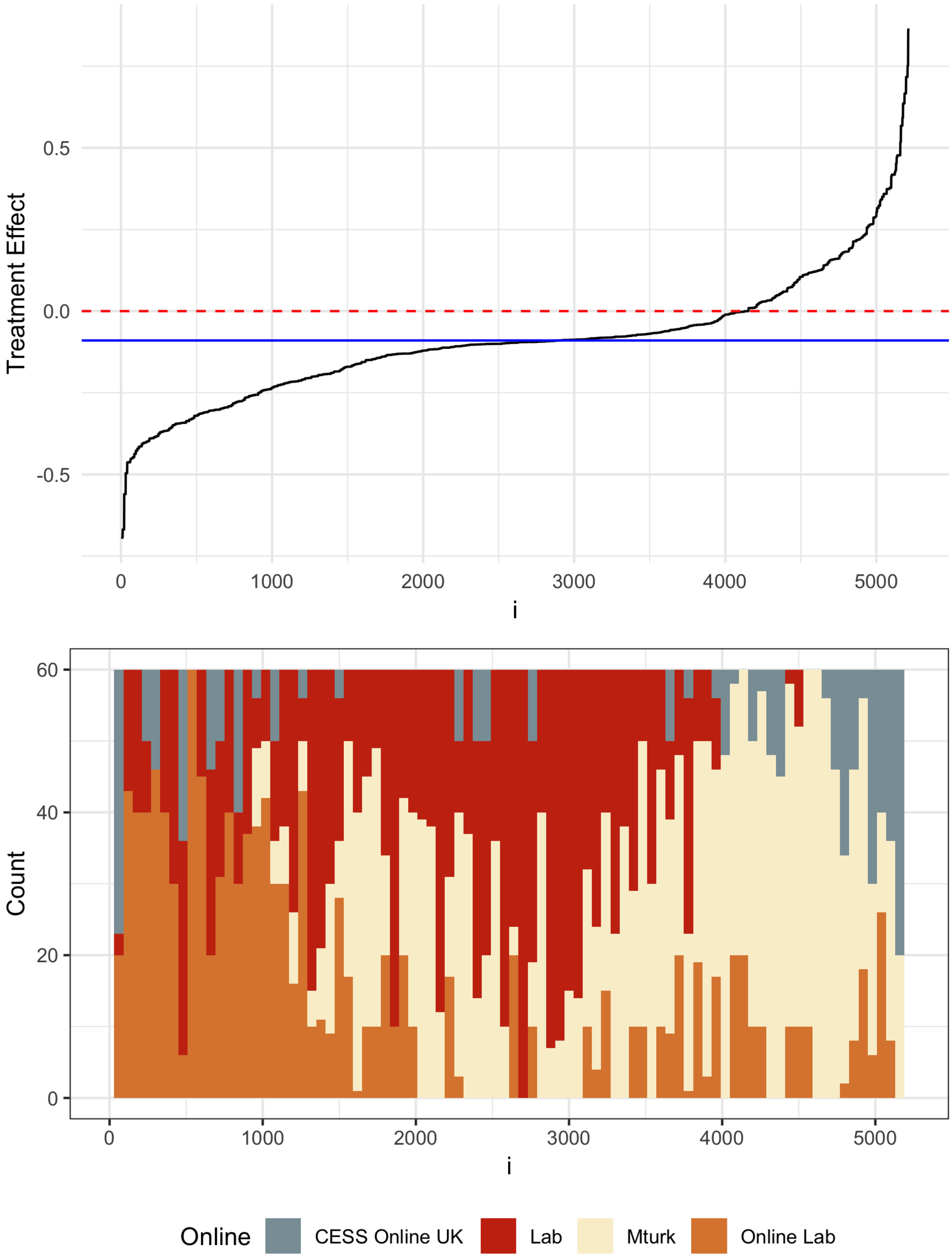
Table A3: Heterogeneous treatment coefficients and interactions using iterated LASSO model

Variable	Coefficient
Treatment	-0.053
MTurk	0.026
Age	0.005
Age^2	-0.000
Gender	0.001
Ability	-0.297
Ability^2	-0.117
Treatment × MTurk	0.129
Treatment × CESS Online UK	0.314
Treatment × Online Lab	0.090
Treatment × Ability	-0.016
Treatment × Age	-0.001
Treatment × Gender	0.029
Online Lab × Ability	0.392
Online Lab × Gender	0.172
MTurk × Ability	0.283
MTurk × Age	0.000
MTurk × Gender	-0.039
CESS Online UK × Ability	-0.445
CESS Online UK × Age	0.000
Ability × Age	0.000
Ability × Gender	-0.096
Age × Gender	-0.001
Treatment × Online Lab × Ability	-0.290
Treatment × Online Lab × Age	0.016
Treatment × Online Lab × Gender	-0.254
Treatment × MTurk × Ability	0.577
Treatment × MTurk × Gender	-0.017
Treatment × CESS Online UK × Ability	1.73
Treatment × CESS Online UK × Age	-0.004
Treatment × CESS Online UK × Gender	-0.188
Treatment × Ability × Age	-0.019
Treatment × Ability × Gender	0.315
Treatment × Age × Gender	0.008
Treatment × Ability^2	-0.196
Treatment × Age^2	0.000
Intercept	0.578
ATE	-0.090

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 -0.09. 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. Mturk and CESS Online UK subjects' estimated treatment effects are predominantly located towards the right, positive end of the spectrum, although there is some clustering of CESS Online estimated effects towards the left hand side of the spectrum unlike in the BART model.

Figure A5: FindIt estimated heterogeneous effects including covariate interactions



Appendix B Indian Vignette Experiment

Table B4: Subject assignment across modes, error group and treatment

Mode	Error	Incentivised?	Control	Treatment	Total
MTurk	High	No	46	47	93
MTurk	Neutral	No	56	47	103
CESS Online	High	No	57	48	105
CESS Online	Neutral	No	42	53	95
MTurk	High	Yes	64	47	111
MTurk	Neutral	Yes	44	45	89

Table B5: Results of Indian vignette experiment with age and gender controls

	Model					
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	−0.728	−0.682	−3.798***	−3.296***	−1.305**	−0.956***
	(0.477)	(0.477)	(0.511)	(0.494)	(0.513)	(0.331)
Age	0.038	0.049	0.031	0.008	0.036	0.040
	(0.029)	(0.033)	(0.024)	(0.024)	(0.029)	(0.028)
Gender: Male	−0.215	−0.494	0.553	−0.567	−0.202	0.457
	(0.523)	(0.497)	(0.563)	(0.516)	(0.516)	(0.333)
Gender: Other	−1.837					
	(2.458)					
Constant	−67.688	−89.259	−53.185	−8.636	−62.581	−72.020
	(58.674)	(65.382)	(47.330)	(47.828)	(57.578)	(56.612)
Error	Neutral	High	Neutral	High	Neutral	High
Mode	MTurk	MTurk	CESS Online	CESS Online	MTurk	MTurk
Incentivised?	No	No	No	No	Yes	Yes
Observations	103	93	95	105	89	111
Adjusted R^2	0.007	0.037	0.376	0.288	0.045	0.082

Note: *p<0.1; **p<0.05; ***p<0.01