Multi-Modes for Detecting Experimental Measurement Error

Appendix

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

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

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

Pablo Beramendi
Duke University
pablo.beramendi@duke.edu

July 5, 2019

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.[1] 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 Rate	Tax Rate	Report 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;

---

[1] All replication materials are available from the Political Analysis Dataverse (Duch et al., 2019).

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

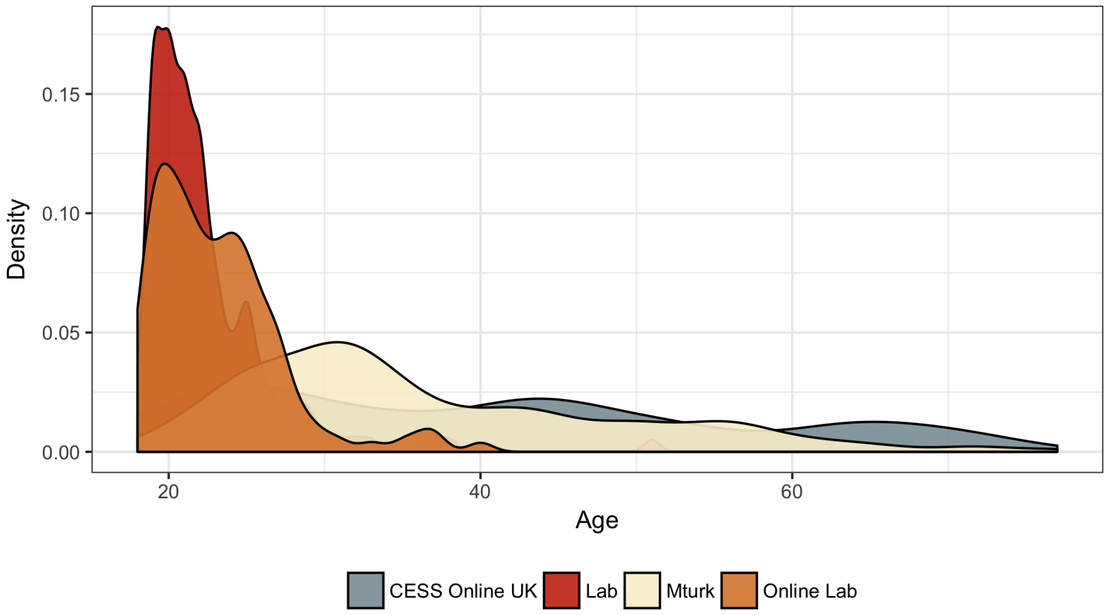


Figure A1: Age distribution of subjects

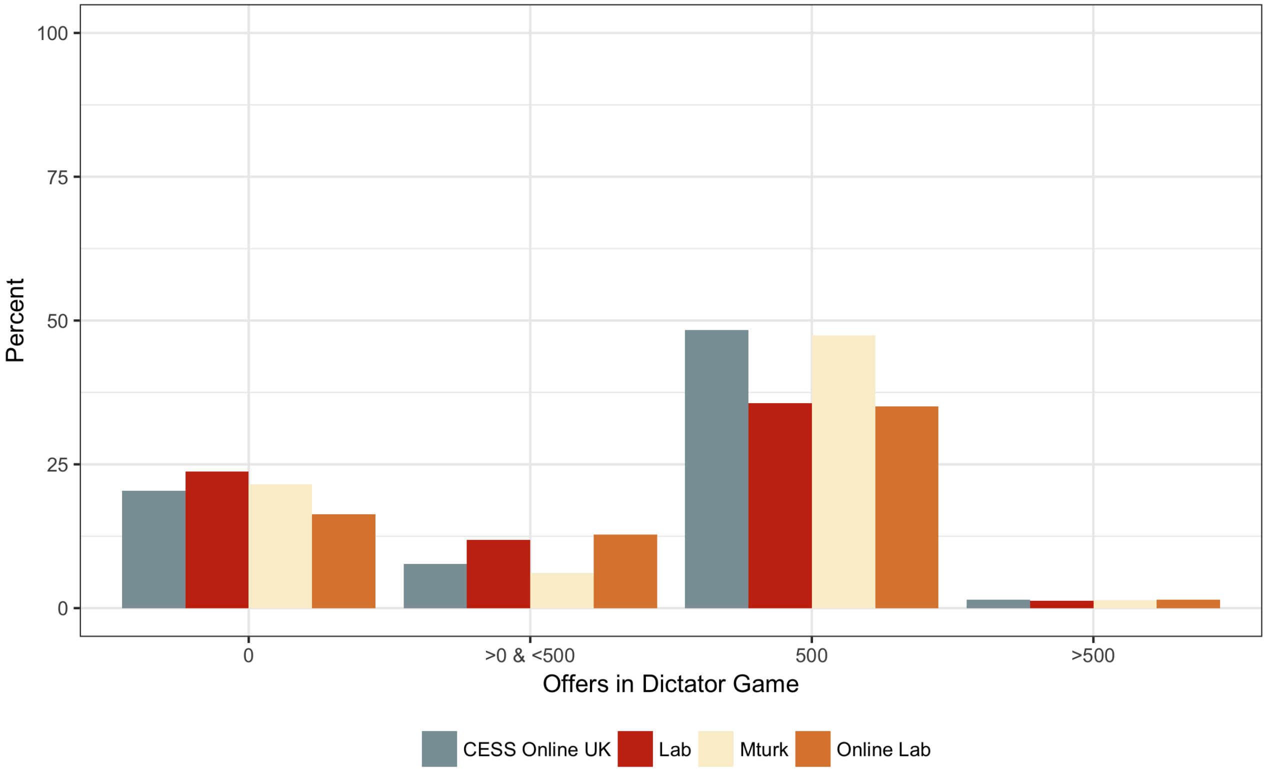
we would expect, the undergraduate student subjects both in the lab and online are even younger on average. The general UK online panel subjects are similar to MTurk subjects. The age distributions for MTurk and UK online are significantly different from UK lab and online in both t-test and Wilcoxon 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

Figure A2: Dictator Game



describes the allocation of ECUs to the recipients dividing the subjects into those that gave nothing to the other person, gave something but less than half, those that split the ECUs evenly and those that gave more than half. A large proportion of subjects either allocate nothing or a half of the endowment to the recipients. The average amount allocated to the recipient is 286 by students in the lab, 303 by students online, 329 by the general UK panel and 307 by Mturk workers.

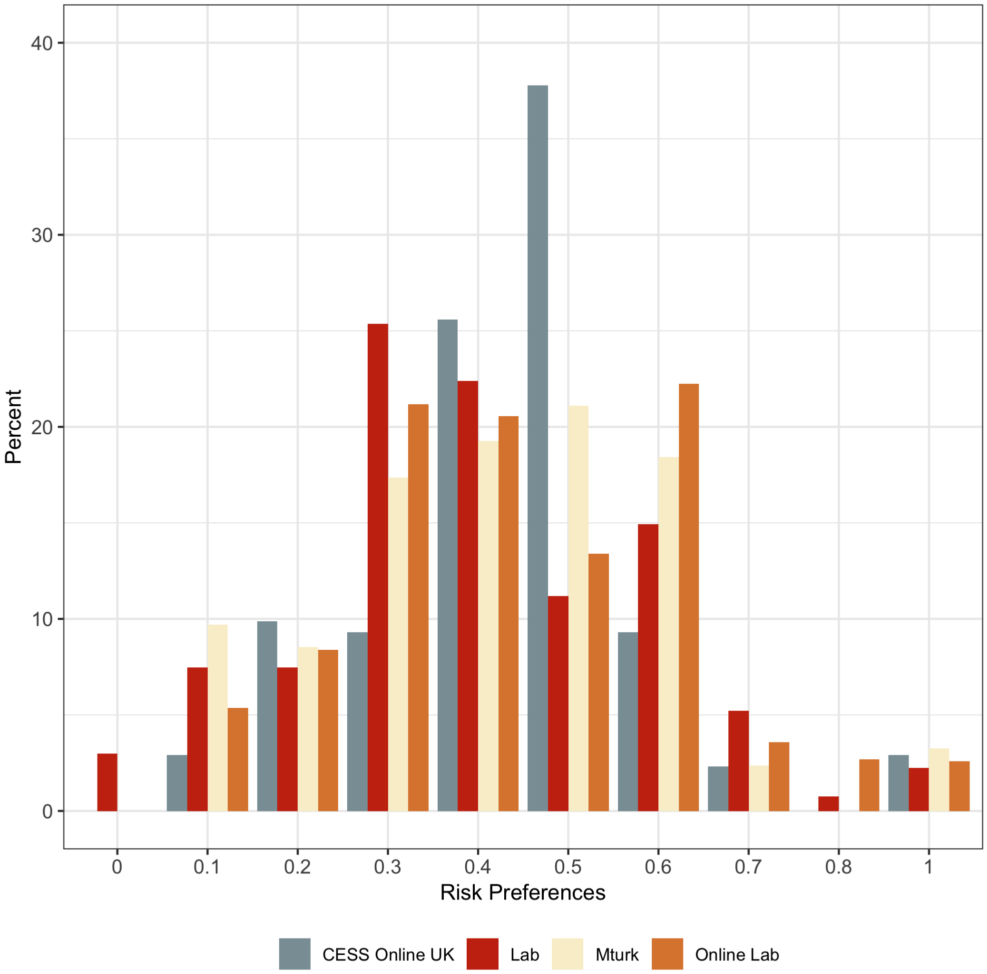
Students are more likely to offer nothing when they are in the lab, but in both t-test and Wilcox rank sum tests, the difference between students in the lab and online is insignificant. In contrast, the UK Online panel and Mturk subjects are significantly more generous than the two student subject pools.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

Figure A3: Risk Preference



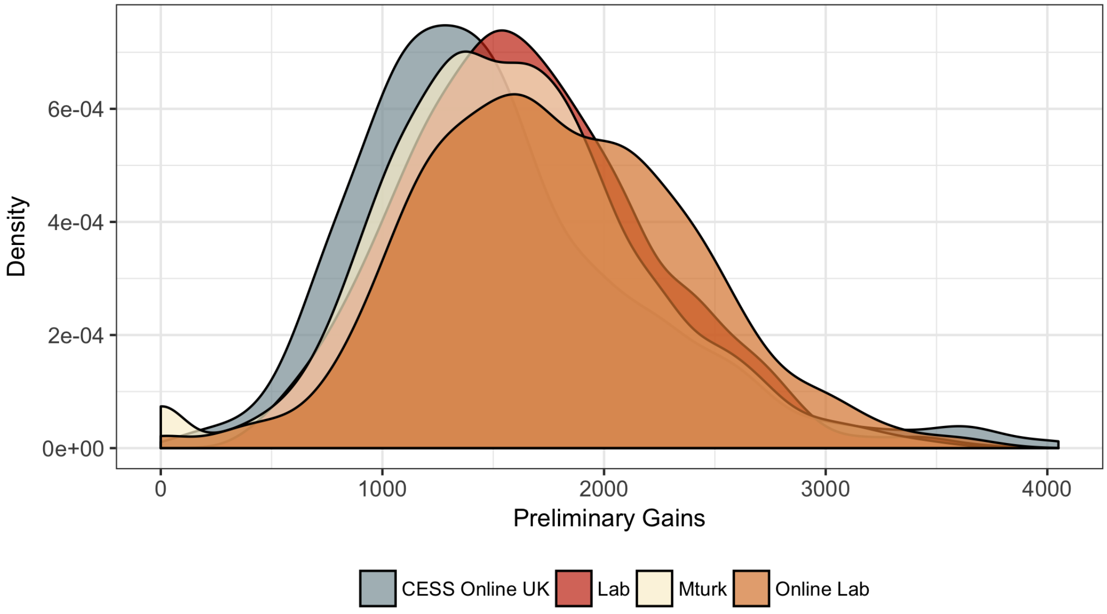
from the same lab subject pool playing the game online and in the lab does not reach conventional levels of significance ($p > 0.1$). And at least two of the three different online subject pools made very similar average offers in the Dictator Game. On the second incentivized risk preference experiment 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

	Wild				PCB			
	Lab	Online Lab	Online UK	MTurk	Lab	Online Lab	Online UK	MTurk
Constant	0.00	0.07	0.00	0.00	0.00	0.07	0.00	0.00
Ability Rank	0.00	0.21	0.46	0.22	0.00	0.21	0.44	0.21
20% Deduction	0.11				0.12			
30% Deduction	0.12	0.01	0.73	0.76	0.12	0.01	0.72	0.77
No Audit	0.00	0.07	0.11	0.84	0.00	0.08	0.13	0.83
Age	0.06	0.48	0.95	0.49	0.07	0.48	0.95	0.49
Gender (1 = Female)	0.98	0.19	0.84	0.95	0.98	0.18	0.85	0.95

Table A2: Wild and PCB clustered p-values

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).[2] 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.[3]

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).

---

[2] 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.

[3] 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.

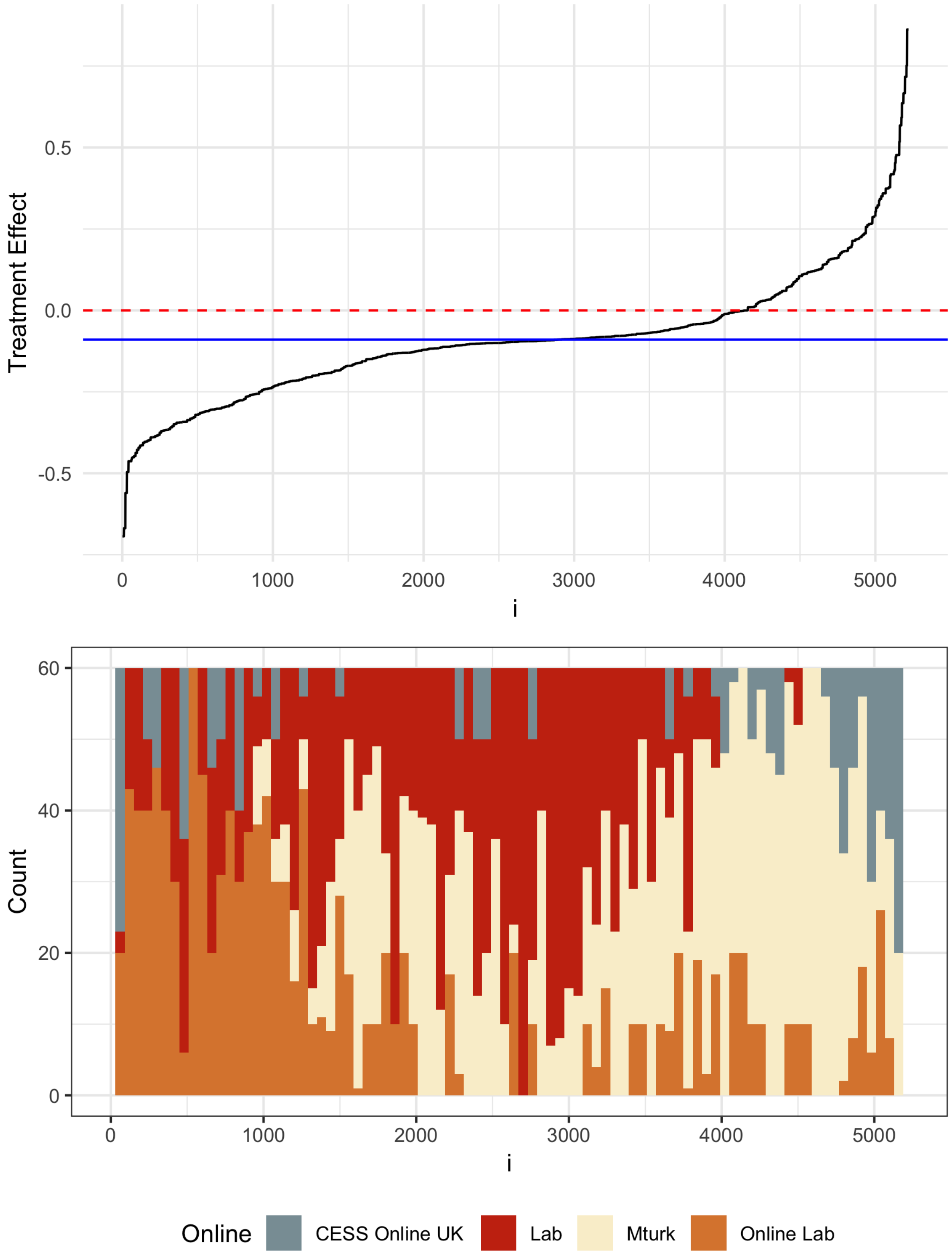
Variable	Coefficient
Treatment	-0.053
MTurk	0.026
Age	0.005
Age^2	-0.000
Gender	0.001
Ability	-0.297
Ability^2	-0.117
Treatment × MTurk	0.129
Treatment × CESS Online UK	0.314
Treatment × Online Lab	0.090
Treatment × Ability	-0.016
Treatment × Age	-0.001
Treatment × Gender	0.029
Online Lab × Ability	0.392
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

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

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

Mode	Error Manipulation	Incentivised?	Authentic News (N)	Fake News (N)
MTurk	Control	No	56	47
MTurk	High	No	46	47
CESS Online	Control	No	42	53
CESS Online	High	No	57	48
MTurk	Control	Yes	44	45
MTurk	High	Yes	64	47

Table B4: Number of participants in each mode and manipulation arm combination

	Dependent variable:					
	Model					
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	−0.73	−0.68	−3.80	−3.30	−1.30	−0.96
	(0.48)	(0.48)	(0.51)	(0.49)	(0.51)	(0.33)
Age	0.04	0.05	0.03	0.01	0.04	0.04
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)
Gender: Male	−0.21	−0.49	0.55	−0.57	−0.20	0.46
	(0.52)	(0.50)	(0.56)	(0.52)	(0.52)	(0.33)
Gender: Other	−1.84					
	(2.46)					
Constant	−67.69	−89.26	−53.18	−8.64	−62.58	−72.02
	(58.67)	(65.38)	(47.33)	(47.83)	(57.58)	(56.61)
Error Arm	Control	High	Control	High	Control	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	0.01	0.04	0.38	0.29	0.05	0.08

Table B5: Regression results for Indian vignette experiment with age and gender controls

	Model	
	(1)	(2)
Treat	-2.16	-1.70
	(0.78)	(0.57)
Constant	8.52	8.47
	(0.54)	(0.37)
Error Arm	Control	High
Mode	MTurk	MTurk
Incentivised?	Yes	Yes
Observations	52	54
R^2	0.13	0.14
Adjusted R^2	0.11	0.13
Residual Std. Error	2.82 (df = 50)	2.07 (df = 52)
F Statistic	7.59 (df = 1; 50)	8.79 (df = 1; 52)

Table B6: Regression results for high attention subjects in attention-incentivised arm (where participant identified *only* the Indian Electoral Commission).

References

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.

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.

Duch, Raymond, Denise Laroze, Thomas Robinson and Pablo Beramendi. 2019. "Replication Data for: Multi-Modes for Detecting Experimental Measurement Error.".
URL: *https://doi.org/10.7910/DVN/F0GMX1*

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

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

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

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

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.