ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Dickinson, David L.; McEvoy, David M.

Working Paper

Further from the Truth: The Impact of In-Person, Online, and mTurk on Dishonest Behavior

IZA Discussion Papers, No. 13686

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Dickinson, David L.; McEvoy, David M. (2020) : Further from the Truth: The Impact of In-Person, Online, and mTurk on Dishonest Behavior, IZA Discussion Papers, No. 13686, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/227213

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13686

Further from the Truth: The Impact of In-Person, Online, and mTurk on Dishonest Behavior

David L. Dickinson
David M. McEvoy

SEPTEMBER 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13686

Further from the Truth: The Impact of In-Person, Online, and mTurk on Dishonest Behavior

David L. Dickinson
Appalachian State University, IZA and Chapman University

David M. McEvoy
Appalachian State University

SEPTEMBER 2020

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Further from the Truth: The Impact of In-Person, Online, and mTurk on Dishonest Behavior*

Recent policies require some interactions previously conducted in close social proximity (e.g., school, workplace) to take place remotely, which motivates our investigation of how in-person versus online environments impact honesty. We modify a well-known coin-flip task and examine the influence of going from the physical laboratory environment, to online with identifiable participants (same lab subject pool), to online with anonymous participants using mTurk. Surprisingly, while a simple move from in-lab to online (using the same subject pool) appears to increase "fake effort" – those who likely never flip the coin - it does not predict more dishonest behavior when there is a monetary incentive to cheat. The most socially distant and anonymous participants (mTurk) are more likely to be deemed cheaters in our analysis—these individuals report coin flip outcomes consistent with cheating for monetary gain. Implications of our findings indicate the greatest risk of potentially costly dishonest behavior results when anonymity, not just social distance, is high.

JEL Classification: C91, D90

Keywords: social distance, cheating, coin flip, anonymity, behavioral economics, experiment

Corresponding author:
David L. Dickinson
Economics Department
Appalachian State University
Boone, NC 28608
USA

E-mail: dickinsondl@appstate.edu

* We thank a Deans Club grant and the Department of Economics at Appalachian State University for funding these experiments. Thanks to Jason Aimone, Dave Bruner, and participants in the socially distant ASU virtual (Zoom) brownbag for helpful comments on early drafts of this paper.

1. INTRODUCTION

Understanding motivations to cheat is important for the design of effective policies, including encouraging tax compliance, adherence to environmental regulations and deterring political fraud. How factors like online or anonymous decision making impact dishonesty is more important than ever given the recent social distancing mandates to help control the COVID-19 virus. There may be, for example, an unintended consequence that dishonest behavior increases with a move to an online environment (e.g., cheating in remote learning environments, inflating self-reported effort or hours worked from home). This paper examines outcomes in an honesty task both in-lab and in an online environment. The design, hypotheses, and analysis plan were all preregistered on the Open Science Framework.[1] In both lab and online environments, we also implemented treatments with and without a monetary temptation for dishonesty. We provide evidence that not only is dishonesty more likely when monetary temptation exists, but dishonest behavior is exacerbated in the most socially distant *and* anonymous environments.

In economics, one approach to investigate cheating behavior is the "coin flip" or "die rolling" task (e.g., Houser et al, 2012; Fischbacher and Föllmi-Heusi, 2013; Gächter and Schultz, 2016;).[2] In these tasks, participants are asked to flip a coin (or roll a die), and they are paid based on the reported outcome. Because the outcome is private and known only to the subject, experimentalists use the statistical properties of a fair coin flip (e.g., 50% of the flips are expected to be Heads) to identify cheating at the group level—with payoff incentive to report Heads, a sample with significantly more than 50% Heads, or die roll outcomes not uniformly distributed across all possibilities, is deemed to be cheating. Using such paradigms, differences in dishonesty across countries and societies is positively correlated with rule-of-law violations (Gächter, S. and Schulz, 2016) and negatively correlated with GDP (Hugh-Jones, 2016). Other research has shown a direct, within-subjects, link between lab-based cheating and dishonesty in a field setting (Potters and Stoop, 2016), which suggests practical relevance of lab-based research on dishonesty. Cheating appears more likely when one can justify the behavior in some way (e.g, Houser et al, 2012), or when the dishonesty harms others in a more abstract way (e.g., the researcher's budget) rather than other participants in the experiment (Köbis et al, 2019).

A recent conclusion from the meta-study in Abeler et al (2019) was that preferences for honesty and *appearing* honest are main determinants of honest behaviors.[3] This is consistent with the theme of our results. Because our lab vs online comparison is one where there is an increase in social distance between the experimenter and the decision maker, one might predict an increase in dishonesty when there is a monetary incentive to cheat (i.e., temptation). Others have also suggested that lesser monitoring or more anonymous/socially-distant interactions increase dishonesty (Rosenbaum et al, 2014; Cohn et al, 2018). Our experimental design allows us to examine the impact of somewhat varied

---

[1] See preregistration plans at osf.io/psztk (Dickinson and McEvoy, 2019) and osf.io/wa4kr (Dickinson and McEvoy, 2020)

[2] Other tasks examine deception that harms the payoff of another experiment participant (e.g., the Sender-Receiver game of Gneezy, 2005). Such environments may differ in important ways from environments in which dishonesty only affects the decision maker (or an abstract researcher budget. See meta-analysis in Köbis et al, 2019). Capraro (2017) is an example where dishonesty that would impact another participant was reduced with time-pressure.

[3] Others have not found robust gender differences in individual versus group decision-making tasks (Muehlheusser et al, 2015, Ezquerra et al, 2018). Another line of research has examined whether dishonesty results from deliberative or automatic thinking, often inferring deliberation from longer response times (Rubinstein, 2007; Kahneman, 2011). Initial conclusions were that honesty required deliberation, and therefore time (Shalvi et al, 2012; Bereby-Meyer and Shalvi, 2015). However, that viewpoint has been challenged recently (Krajbich et al, 2015).

degrees of social distance and anonymity on dishonest behavior. In one set of treatments, subjects were recruited from our university and participate in an in-person lab-based experiment with close interactions between an individual participant, her peer participants, and the experimenter—this took place in the Fall of 2019 prior to COVID 19 policy implementation. The next set of treatments used a similar subject pool to the lab-based experiments but the sessions were conducted online to increase the social distance between all parties. Such treatments do not make the participant totally anonymous, however, as the experimenter has much of the same identifying information about the participants from the database. The final set of treatments utilized the more fully anonymous online mTurk workforce as a subject pool. This set of treatments produces the largest increase in social distance as the participants do not interact with the experimenter or their peers, and the participants remain anonymous.

Our study is timely as we aim to better understand the potential unintended consequences on ethical decision making of face-to-face versus online decision making, which can also help improve our understanding of the impacts of widely mandated social distancing measures in response to the COVID 19 pandemic. Such measures have required instructors to move in-class teaching to remote (online) learning environments, many occupations have transitioned to remote work arrangements, and behavioral researchers have moved to 3rd party platforms for recruiting participants.

2. METHODOLOGY

Participants were recruited for a short 10-15 minute experiment. For some treatments, we recruited from the authors' experimental economics subject pool using the Online Recruitment System for Economic Experiments, or "ORSEE" (Greiner, 2015). In other online treatments, we recruited participants using Amazon's Mechanical Turk (mTurk). The experiment was approved by the Institutional Review Board at the investigators' institution. The survey elicited age, gender, minority status, a 6-item Cognitive Reflection Task (CRT) (Primi et al, 2016) that extends the original CRT (Frederick, 2005), and it contained an attention-check "poison pill" question prior to the coin flip task (See Appendix B). In our variation of the coin-flip task we elicited the number of HEADS resulting from a series of 10 coin flips. For the in-lab task, we provided a quarter at each individual lab station. For the online treatments, a survey page prior to the decision task asked individuals to *find a quarter/coin* that could be flipped and something to write with *prior* to moving on to the next page.[4] It is on that next page where task incentives were described, participants were asked to flip the coin 10 times (making note of the order) and record the outcome. In addition to eliciting the overall outcome of the 10 coin flips, we also recorded the time spent on that flip-and-decision page as the decision task response time. Finally, one additional modification to the standard coin-flipping task is that a follow-up page asked participants to record the flip sequence. This additional outcome measure provides us with another method to statistically examine honesty, as some may report highly unlikely flip sequences.[5] All three outcome measures and hypotheses were preregistered for this study.

Our design systematically varied each of two factors across six treatments.[6] In the *Fixed* pay treatments, subjects were instructed that their payoff did not depend on the outcome of the coin flips,

---

[4] This is important because it helps maintain consistency in requiring participants to flip a coin. It is, of course, harder to verify in the online treatments whether or not a coin was flipped. For this reason, our data on coin flip response times is important, but in all cases it is possible that subjects did not flip the coin as requested.

[5] We thank Jason Aimone for this helpful suggestion to elicit the flip sequence.

[6] See preregistration plans at osf.io/psztk and osf.io/wa4kr. The initial pre-registration was completed *prior* to any data collection, and focused on running in-lab and mTurk participant treatments under conditions of Fixed and Variable payments. The second registration was added to introduce additional online treatments using the same

whereas in *Incentives* treatments there was the temptation to cheat—subjects earned an additional payment of $0.25 for each reported HEADS that was in addition to the promised fixed payment for participation. We also varied the level of social distance and anonymity of the participants in the task. We conducted treatments in-lab using participants recruited from the ORSEE database. The *Lab* participants came to the (physical) experimental economics laboratory at our institution for the short experiment. Participants from the ORSEE database were also recruited for online treatments that specified the participants would be sent a survey link and should not come to the physical lab. Finally, the mTurk participants were recruited via standard practices through Amazon's Mechanical Turk.[7] Payments were made in cash to in-lab participants, by Amazon gift code to ORSEE participants recruited for the online treatments, or through the standard *mTurk* payment process that promised either a *Fixed* payment or the *Incentive* pay, which we paid using the mTurk bonus payment feature.[8] As such our six total treatments are *ORSEE-Lab Fixed, ORSEE-Lab Incentive, ORSEE-online Fixed, ORSEE-online Incentive, mTurk Fixed,* and *mTurk Incentive*.

We recruited to achieve a sample size of 100 participants in each treatment. In addition to our hypothesis that incentives matter (i.e., *Incentives* creates temptation), we also hypothesized that more statistically unlikely reports of HEADS would be observed in treatments that were online and anonymous (i.e., *ORSEE-online* and *mTurk*, respectively, compared to the *ORSEE-Lab* samples) when there was a temptation to be dishonest. Given the mixed results in the literature regarding response time, deliberation, and honesty, our only hypothesis regarding response times was that the *mTurk* and *ORSEE-online* samples would have faster response times as they would be less likely to actually flip a coin given the complete lack of observability online and out-of-lab. Our pre-registered hypotheses are:

H1: Statistical cheating is more prevalent online and when anonymity is increased. The predicted ordering of statistical cheating is: *mTurk* > *ORSEE-online* > *ORSEE-Lab*

H2: Statistical cheating is more prevalent when the marginal incentive to cheat is positive (i.e., more dishonesty will result from *Incentive* compared to *Fixed* payments).

H3: Response times (RT) will be faster when social distance (i.e., online) and anonymity are increased. The predicted ordering of RT is: *mTurk* < *ORSEE-online* < *ORSEE-Lab*

---

subject database as the in-lab treatments. Thus, this second registration happened after data analysis of the original 4 treatments, but *prior* to data collection for the additional treatments.

[7] The in-lab sessions were conducted in October-November, 2019, and so did not violate any social distancing mandates. The *ORSEE*-online sessions were conducted in April, 2020, where it was made clear that participants would *not* come to the physical lab (plus, most of the database students were no longer in town given the University had moved to remote classes *and* required on-campus student to vacate their dorm rooms). MTurk workers conditions were: U.S. resident, 18 or older, and had a ≥ 95% completion rating on their previous mTurk assignments.

[8] Participants from the ORSEE lab database were paid $6.00 in the Fixed treatment and $5.00 plus $0.25 per HEAD in the Incentive treatment (expected earnings of $6.25), whether in-lab or online for consistency of pay rates for those in our database. Online *mTurk* participants were paid a lesser hourly pay rate to be more in line with *mTurk* compensation expectations. *mTurk o*nline subjects were paid $2.00 for the Fixed treatment and $1.00 plus $0.25 per HEAD in the Incentive treatment (expected earnings of $2.25), which still exceeds the *mTurk* reservation wage by a large amount. The average time for the lab sessions was roughly 15 minutes and the average time for the online sessions was roughly 10 minutes.

H4: Statistically less likely sequences of coin flips will be reported when the incentive to cheat is positive (*Incentive* treatments), and when social distance or anonymity are increased. The predicted ordering of unlikely coin flip reports is: *mTurk* > *ORSEE-online* > *ORSEE-Lab*

Though not formalized as a numbered hypothesis, we also preregistered our plan to combine information from all three outcome measures (# Heads, RT, flip sequence) to create and indicator variable that identifies an individual as a likely cheater. In the *Incentives* treatments, we therefore also predict that the likelihood a subject is identified as a cheater should be ordered as follows across treatments: *mTurk* > *ORSEE-online* > *ORSEE-Lab*.

3. RESULTS

In each of our treatments, our final sample size depends on whether the participant passed the poison pill question in the survey. Our complete sample consists of: *ORSEE-Lab Fixed* n=103; *ORSEE-Lab Incentive* n=102; *ORSEE-online Fixed* n=104; *ORSEE-online Incentive* n=113; *mTurk Fixed* n=111; *mTurk Incentive* n=121. Conditioning on passing the poison pill question (the "PPP" subsample), these respective sample sizes are 97, 95, 98, 109, 105, and 111.

We first evaluate the data on the coin flip outcomes, HEADS, to test hypotheses H1 and H2. Panel A of Figure 1 shows the cumulative distribution function of HEADS outcomes in our six treatments, along with the expected distribution that would result from a fair coin set of 10 flips. The average number of HEADS reported in the *Fixed* payoff treatment for the *ORSEE-Lab, ORSEE-online,* and *mTurk* samples was 5.13, 4.67, and 5.16, respectively, while the respective values in the *Incentives* treatment were 5.66, 5.28, and 6.15. For the PPP subsample, these averages were 5.14, 4.70, and 5.21 in the *Fixed* treatment, and 5.69, 5.31, and 6.50 for the *ORSEE-Lab, ORSEE-online,* and *mTurk* samples, respectively. As pre-specified in our analysis plan, one-sided p-values will be reported for our directional hypotheses above, and we will report results from appropriate non-parametric tests and conditioned multivariate analysis. We report nonparametric results on the subset of participants who passed the poison pill (the "PPP" subsample), although for transparency and comparability, we report multi-variate analysis on the PPP subset as well as the full data set. Hypotheses H1 and H2 can first be examined using Mann-Whitney tests of medians.

3.1 The effect of social distance and anonymity

Recall, our Hypothesis 1 (H1) was that statistical cheating will increase in the online environment and with increased anonymity. Using data *only* from the *Incentive* pay treatments where there exists a temptation to be dishonest, we test the HEADS reports across the following pairs of treatments: *ORSEE-Lab* vs *ORSEE-online*, *ORSEE-Lab* vs *mTurk*, *ORSEE-Lab* vs *mTurk*. Mann-Whitney tests for each pair-comparison in the *Incentive* pay treatment result in: *ORSEE-Lab* vs *ORSEE-online* ($z = 1.926$, $p = .054$), *ORSEE-online* vs *mTurk* ($z = -4.571$, $p < .01$), *ORSEE-Lab* vs *mTurk* ($z = -2.910$, $p < .01$). Results for two of the three treatment comparisons are consistent with H1. Surprisingly, the median number of HEADS reported in *ORSEE-Lab* is *greater* than in *ORSEE-online*, which is contrary to H1 regarding the impact of social distance on dishonestly. A separate z-test shows that both *ORSEE* samples in the *Incentive* treatments reveal statistical evidence of cheating in the PPP subsample data with an average HEADS reports greater than 5 of 10 flips (mean HEADS=5.31 in *ORSEE-online Incentives*, mean HEADS=5.69 in *ORSEE-Lab Incentives*: $p < .01$ for both the one-sample z-tests). This is also true of the *mTurk* sample

data in *Incentives* with mean HEADS report of 6.50 ($p < .01$). This initial unconditional analysis supports the hypothesis of cheating in the all *Incentive* treatments. But, while the highest number of HEADS is reported when temptation is present among *mTurk* participants, the lowest level of statistical dishonesty is found in the *ORSEE-online* sample.

The multivariate analysis to test H1 is found in Appendix Table A1 and summarized in the coefficient plot Figure 2. Specifications estimated on both the full sample and the PPP subsample include treatment controls only, the addition of pre-registered covariates, and the inclusion of the necessary interaction term of the *Incentives* treatment with the *ORSEE-online* and with *mTurk* participant groups to properly test H1. The fully specified model is contained in equation (1) below.

$$\begin{aligned} NumHeads &= \beta_0 + \beta_1 ORSEEonline + \beta_2 mTurk + \beta_3 Incentives + \\ &\beta_4 ORSEEonline * Incentives + \beta_5 mTurk * Incentives + \gamma Controls + \mu \end{aligned} \tag{1}$$

Focusing exclusively on *Incentives* treatments, H1 suggests that the number of heads in the *ORSEE-online* and *mTurk* treatments should be higher than the *ORSEE-Lab* treatment, and the ranking should be *mTurk* > *ORSEE-online* > *ORSEE-Lab*. The change in predicted HEADS from *ORSEE-Lab* to *ORSEE-online* is $\beta_1 + \beta_4$ (i.e., combined large-square symbol coefficients in Figure 2 for these two variables). An F-test shows that this combined effect is negative (i.e., the online environment reduces the number of heads reported) and statistically significant ($p = .045$). This is consistent with the initial nonparametric test result, but it is surprising and opposite to our preregistered hypothesis. The change in predicted HEADS from *ORSEE-Lab* to *mTurk* treatment is $\beta_2 + \beta_5$, which is positive and highly significant ($p < .01$) suggesting more cheating in *mTurk* relative to *ORSEE-Lab*. Finally, it is clear that the *mTurk* treatment has a positive and significant effect on HEADS compared to the *ORSEE-online* treatment ($p < .01$). These results are similar if restricting analysis to the PPP sample, with the exception that the surprisingly fewer number of HEADS reported in *ORSEE-online* relative to *ORSEE-Lab* drops in significance to a marginal level ($p = .056$). In other words, our findings suggest that when the temptation for dishonesty is present, an increase in social distance by moving to an online environment does not, by itself, increase HEADS report dishonesty beyond what one would expect in the less socially distant environment of the physical laboratory experiment. However, increasing the level of anonymity via *mTurk* increases dishonesty in HEADS reports by a larger extent than we estimated in either the *ORSEE-Lab* or *ORSEE-online* participants.[9]

3.2 The effect of monetary incentives

This hypothesis (H2) compares the number of HEADS reported in *Fixed* versus *Incentive* treatments for a given participant pool (*ORSEE-Lab, ORSEE-online,* or *mTurk*). Mann-Whitney test results clearly support H2 for each comparison: *ORSEE-Lab* ($z = -2.170$, $p < .015$); *ORSEE-online* ($z = -2.881$, $p < .01$; *mTurk* ($z = -5.126$, $p < .01$). Similar results are found when pooling all subjects for a test of HEADS reported in *Incentives* (n=315 total) compared to *Fixed* (n=300 total) treatments ($z=-5.778$, $p < .01$). Thus far, the analysis clearly supports the hypothesis that temptation increases dishonesty and that *mTurk* participants are the most dishonest across our participant pools.[10] Hypothesis 2 is also clearly supported

[9] As an exploratory analysis, we found no robust effects of individual-specific characteristics on HEADS reports.
[10] Though *ORSEE-online* participants reported unexpectedly *fewer* HEADS compared to *ORSEE-Lab* participants, *adding* the temptation (*Incentives*) induced significantly more HEADS reports in *both ORSEE-online* and *ORSEE-Lab*

by the results in Figure 2 (see Appendix Table A2), as indicated by the statistically significant and positive main effect of *Incentives* in increasing the number of HEADS reported, in general ($p < .01$ in all models), and the lack of statistically significant *negative* coefficient estimates on either interaction term.

3.3 Examining response times

We next turn to an examination of Hypothesis 3 (H3) regarding response times (RT), the cumulative distributions of which are shown in panel B of Figure 1. The prediction is that coin-flip response times will be ordered from slowest to fastest in *ORSEE-Lab* (n=192), *ORSEE-online* (n=201), and then *mTurk* (n= 216). Mann-Whitney results show that RTs are not significantly different in *ORSEE-online* compared to *ORSEE-Lab* ($z = -.480$, $p > .10$), but RTs are faster in *mTurk* compared to *ORSEE-online* ($z=6.579$, $p < .01$), and compared to *ORSEE-Lab* ($z=7.577$, $p < .01$). The regression analysis in Table 1 confirms these nonparametric findings. Across all models in Table 1 (and, in both full and PPP subsamples) we find that RTs are faster in *mTurk* than *ORSEE-Lab* and *ORSEE-online* ($p < .01$ in all models). However, there are no significant differences in RTs between *ORSEE-online* and *ORSEE*-Lab ($p > .10$ in all instances). Thus, we find partial support for H3 as RTs only decrease as a main effect in *mTurk* (i.e., moving to the online environment *and* increasing anonymity). While we did not preregister a hypothesis regarding the impact of *Incentives* on RTs, an exploratory finding is that RTs are consistently greater when temptation is present.[11] Other exploratory findings are that *Minority* participants have slightly longer RTs, and in the full sample CRT score may predict RTs.[12]

3.4 Examining the sequence of coin flips

Finally, our Hypothesis 4 (H4) is that more unlikely sequences of reported coin flips will occur when monetary incentives to cheat are present or in online environments where we hypothesized a decreased likelihood of actually flipping a coin. To test H4 regarding the likelihood of the reported flip sequence, we first conducted runs tests on each participant's reported flip sequence (n=10 for each).[13] Additionally, reports of the most unlikely outcome (10 HEADS) will be uninformative here because the test compares the observed runs to the expected runs, *given the reported outcome*, and so we discard

---

(H2). Also, these two treatments were administered during the same calendar week, which helps control for unobservable factors specific to treatment timing during the COVID 19 stay-at-home orders.

[11] It would be unclear from our data whether this results from ethical conflict that leads to more deliberation and longer RT, or it could simply be that the more complicated instructions describing the incentive pay added the average 30 seconds (approximately) to all RTs.

[12] We also did not preregister a hypothesis regarding the RT variance, which appears different across treatments and groups (Fig.1, panel B). As an exploratory analysis, we conducted inter-quantile (IQ) regressions with bootstrapped standard errors (1000 replications) on the simple specification with RT as the dependent variable, and controls for *Incentives, ORSEE-online, mTurk* and demographics. The IQ results showed that *Incentives*, *ORSEE-online*, and *mTurk* all significantly increase the RT variance compared to the baseline *ORSEE-Lab Fixed* treatment group ($p < .01$ in all instances).

[13] The normal approximation is appropriate for the binomial test under certain conditions. In our case, the sample (n=10 per subject) and success probability, p, are fixed such that $n \times p = n \times (1-p) = 5$. Some sources view this as the acceptable threshold for using the Z distribution to calculate probabilities, but others set a threshold at 10 (or higher), which would require a ≥ 20-flip sample to use a normal approximation. Because this is just one of our outcome measures, we use the Z distribution in our case.

instances of HEADS =10 reports.[14] It may be worth noting here that all 23 instances of HEADS=9 or 10 in our four samples came from the *mTurk* participant pool, and 19 of those (almost 80%) were in *mTurk Incentives*.

From each runs test we produce a z-score. Next, we identify the *p*-value associated with that z-score, which indicates the likelihood of falsely rejecting a true null hypothesis. The higher the z-score (i.e., the lower the *p*-value), the more statistically unlikely it is to observe the reported HEADS sequence. We next conducted estimations shown in Table 2 that use this *p*-value as the dependent variable in a set of regressions with the same covariates as previously used in Table 1 and Appendix Table A1. As can be seen in Table 2, the estimated coefficient on *Incentives* is not statistically significant in any of the specifications. We only find some evidence of less likely coin flip sequences in *mTurk*. However, in the PPP subsample, the result is only marginally significant ($p < .10$) across specifications. It is perhaps not surprising to find somewhat stronger evidence in the full sample, because it includes those who failed the attention check question and therefore may be less inclined to actually flip a coin (or even know what the task was asking, in the event they failed to read carefully). Overall, the support is weak for H4, although some evidence indicates *mTurk* participants report statistically less likely flips, and low statistical power may be an issue.[15]

3.5 Exploring likely cheaters

Finally, we aimed to incorporate information from our multiple outcome measures in identifying likely cheaters (or, dishonesty), which we hypothesized to be ordered as: *ORSEE-Lab* < *ORSEE-online* < *mTurk* (and greater likelihood of cheating in *Incentives* relative to *Fixed*. Though testing the relationships was not identified as a separate numbered hypothesis, we pre-registered our plan for construction of such cheating identifiers and the testing of these hypothesized relationships.

Heads-cheaters. In creating an indicator variable to identify *Likely Cheaters*, we started with identifying those who reported HEADS > 7 (the cumulative probability of actually flipping 8, 9, or 10 heads on a set of 10 fair coin flips is 5.469%). We can refer to this group as *HEADS-cheaters*. The percentage of *HEADS-cheaters* =1 in the *Fixed* treatments of *ORSEE-Lab, ORSEE-online,* and *mTurk* are 7.22%, 4.08%, and 3.81%, respectively. In the *Incentives* treatments, the respective percentages are 12.63%, 4.85%, and 26.13%.

Fake Flippers. Next, we used RT data to score a variable that indicates participants who likely did *not* actually flip a coin as asked. If we assume the *ORSEE-Lab Fixed* data yield a reasonable benchmark RT distribution of individuals who actually flipped a coin 10 times, then abnormally fast RTs can be used to score an indicator for "*Fake Flipper*". We scored *Fake Flipper*=1 for individuals with RT < 45 seconds, which is approximately at the 1% percentile (43.375 seconds) of the RT distribution in the benchmark *ORSEE-Lab Fixed* for the PPP subsample. So essentially, only 1 person out of about 100 participants who

[14] Flipping 0 HEADS is equally as unlikely as flipping 10 HEADS, but we do not observe any reports of 0 HEADS in our samples.

[15]For completeness, we also conducted the analysis of flip sequence likelihood with the data from HEADS=10 participants. In this case, the coefficient estimates on *mTurk* are no longer statistically significant. Inclusion of the flip sequence for the 10-HEADS reports implies the addition of several *mTurk* flip sequences considered *the most probable* (given there is only one sequence that can produce a 10 HEADS outcome), which would dilute our results regarding the more discriminatory flip sequences. Thus, in our Table 2 analysis we chose to focus on the flip sequence reports that were not the only option available to someone who may have not actually flipped the coin.

were in the lab had a RT this fast.[16] By this measure, we identify the percentage of *Fake Flipper* =1 in the *Fixed* treatments of *ORSEE-Lab, ORSEE-online,* and *mTurk* are 1.03%, 16.33%, and 49.52%, respectively. In the *Incentives* treatments, the respective percentages are 2.11%, 6.80%, and 26.13%. While it is possible that some online participants resorted to virtual coin flippers for this task, they were asked in the instructions to flip an *actual* coin and not proceed to the task page of the survey until they had retrieved a coin.

SequenceBS. Lastly, we uses the likelihood of the reported flip sequence to score a variable *SequenceBS*=1 if the individual reported a sequence that did not produce an absolute z-score of 1.645 or higher (i.e., the level for a p = .10 rejection on a 2-tailed test on the null hypothesis of a statistically likely flip sequence). Recall that this variable derives from the runs test, for which we omitted the data from participants report a 10-HEADS sequence. As with the other measures based on HEADS reported and RTs, this metric is not without limitation, but our aim is to consider the set of metrics together in making conclusions about the likelihood of cheating. By this stringent standard, we identify < 5% *SequenceBS* =1 participants across all treatments.

At this point, we can construct the following dependent variables intended to capture the likelihood of being a cheater: *Likely Cheater* = 1 just refers to a *HEADS-cheater* as defined above; *More Likely Cheater* =1 if *Likely Cheater=1* and either *Fake Flipper* =1 or *SequenceBS* =1; *Very Likely Cheater* =1 if *Likely Cheater* =1, *Fake Flipper* =1, and *SequenceBS* =1. As it turns out, only one participant across all our samples was classified as a *Very Likely Cheater* (an *mTurk-Fixed* participant), and so our estimations focus on modeling the predictors of being classified as a *HEADS-cheater,* a *Fake Flipper*, or a *More Likely Cheater* (i.e., both a *HEADS-cheater* and a *Fake Flipper*). Results were qualitatively, quantitatively (in terms of marginal effects we can estimate), and statistically similar using nonlinear Probit or linear probability estimations, and so we used linear probability models for simplicity and estimated each model with the full set of predictors used in the previous analyses for both the full and PPP samples of data. These results are summarized by the coefficient plots in Figure 3 (full results are in Appendix Table A2). Consistent with our previously reported results, we find that *mTurk* participants are the most likely to be categorized as a cheater given our approach. They are more likely to be *Heads-cheaters* compared to *ORSEE-Lab* or *ORSEE-online* participants in the *Incentives* treatments (see the *mTurk*Incentives* marginal effect, p < .01 in both instances), and they are more likely to be *Fake Flippers* than both *ORSEE-Lab* and *ORSEE-online* participants in either the *Fixed* or *Incentives* treatments (p < .01 in all instances).[17] We also find some support in this analysis for the hypothesized ordering of the likelihood of being a *Fake Flipper*, because we find that a *Fake Flipper* is more likely in *ORSEE-online* than *ORSEE-Lab*. Note that *mTurk* participants are even more likely to be deemed *Fake Flippers* than *ORSEE-online* participants, and this is true in both *Incentives* and *Fixed* payment treatments (note x-axis scale differences in Figure 3). Regarding *More Likely Cheaters*, it is only the *mTurk* participants who are more likely to be deemed a *More Likely Cheater*—this is true in the *Fixed* treatment due to the strong *mTurk* effect of unrealistically fast RTs (i.e., fake flipping), and is true in the *Incentives* treatment.[18]

---

[16] These *ORSEE-Lab* participants were physically in the lab, we provided quarters, and we could hear them hitting the tables, see coins flipping, etc., even though we remained at the front of the lab unable to verify outcomes. Using the 1% percentile of RTs is, of course, a very conservative threshold to identify fake flippers.

[17] We document that *mTurk* participants are more likely to be *Fake Flippers* by testing that the combined coefficients on the *mTurk+(mTurk*Incentives)* are equal to the combined coefficients on *ORSEE-online*+(*ORSEE-online*Incentives*).

[18] Though exploratory in nature, we also report some evidence that older, minority, female, and high *CRT-score* participants are less likely to be deemed cheaters using our categorizations.

DISCUSSION

Our goal was to examine a modified coin-flip task to further our understanding of dishonesty when moving from an in-person to an online environment, with varying degrees of anonymity. These are features of more socially distant interactions that have increased in frequency with, for example, online learning and remote work arrangements. Our data suggest two conclusions: Dishonesty is more likely when temptation is introduced, and it is also more likely with online decision making *and* increased anonymity in our individual-level task. In general, we fail to find strong support for our preregistered hypothesis that increasing social distance alone (i.e., moving from in-lab to online without much additional anonymity) would generate dishonest behaviors at levels between the in-lab and anonymous online participant groups. The one exception is that we found evidence that the online environment alone increased the likelihood that participants were not actually flipping a coin, compared to the in-lab group, but the increase in *Fake Flipping* was not as extreme as with the online plus anonymous *mTurk* subject pool. It may therefore be that increased social distance alone is sufficient to make agents seek shortcuts perceived to be inconsequential, but not sufficient to make one exploit the social distance for dishonest gain. This was a surprising result that *ORSEE*-online participants did not report more HEADS than *ORSEE-Lab* when monetary incentives to be dishonest were present. Regarding the most socially distant and anonymous participants (those in the *mTurk* treatment), we generally reported outcomes consistent with our pre-registered hypotheses. We found that, while temptation significantly increased the frequency of statistically *unlikely* outcomes among all participant groups, the greatest percentage of those deemed dishonest was found in the *mTurk* sample with temptation present.

Of course, some key limitations must be recognized. First, the special circumstances surrounding the timing of our experiment sessions led to some unavoidable differences between treatments. As noted in our preregistration dates and materials for this study, the in-lab and mTurk treatments were administered prior to COVID 19 restrictions (November, 2019). The addition of our intermediate *ORSEE-online* treatment, however, was after COVID 19 restrictions took hold. Thus, while the *ORSEE-online* treatments were able to be administered without concerns given they took place online, these sessions took place in April 2020 in the midst of COVID 19 stay-at-home orders. It is unclear to what extent the development of the pandemic may or may not have influenced one's proclivity towards dishonesty in our task. Also, there are differences in the *ORSEE-online* and *mTurk* subject pools that go beyond the fact that *mTurk* participants were more anonymous than our *ORSEE-online* participants. While we recognize this uncontrolled variation in the samples, it is difficult to vary nothing more than "anonymity" in a convincing way if using the same subject pool as a starting point. The mTurk participants are, on average, older than the *ORSEE-online* participants (34.97 ± 10.10 versus 20.96 ± 3.14 years of age), but the independent control of age in our estimations revealed little statistical significance attributable to *Age*. If anything, there may be an estimated lower likelihood for older participants to be a *More Likely Cheater* (see Appendix Table A2), but there may be other unobservable differences between mTurk and our ORSEE participants that merit further attention.

Notwithstanding these limitations, implications of our findings with respect to recent policies of social distancing to help reduce the spread of COVID 19 are still worth noting. Our results inform how we might view unintended consequences of social distancing policies. Namely, if social distancing is not coupled with anonymity, then a collateral increase in online student or remote worker dishonesty may not increase substantially. Our results did indicate, however, that simple social distancing (i.e., moving from face-to-face to online environments) may increase the incidence of individuals cyber-loafing or "faking it" in terms of effort put forth or following instructions, which may be considered a form of "time theft" with its own moral implications. Regarding predicted increases in dishonesty for monetary gain, it seems the most important variable is the increased anonymity that may be present in certain specialized socially distant interactions. Finally, additional research is also needed to examine the limits of our findings, in terms of the specialized task and/or payoff levels. For example, temptations may vary in

their severity and so a higher stakes scenario may produce stronger temptation that fewer are able to resist. Our results cannot speak to the overall benefits and costs of social distancing or remote working arrangements, they can only highlight the potential that dishonest behavior may be more pronounced in these environments. These findings could also be helpful for developing procedures to maintain the integrity of remote education, work arrangements or utilizing an online workforce. In particular, the results underscore the potentially costly link between anonymity and dishonest behavior.

REFERENCES

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115-1153.

Bereby-Meyer, Y., & Shalvi, S. (2015). Deliberate honesty. *Current Opinion in Psychology*, *6*, 195-198

Capraro, V. (2017). Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Economics Letters*, *158*, 54-57.

Cohn, A., Gesche, T., & Maréchal, M. A. (2018). Honesty in the digital age. *CESifo Working Paper Series No. 6996.*

Dickinson, D. L., & McEvoy, D. (2019, October 21). Honesty in coin flipping tasks. Retrieved from osf.io/psztk (initial treatments)

Dickinson, D. L., & McEvoy, D. (2020, April 20). Honesty in coin flipping tasks. Retrieved from osf.io/wa4kr (additional treatment set)

Ezquerra, L., Kolev, G. I., & Rodriguez-Lara, I. (2018). Gender differences in cheating: Loss vs. gain framing. *Economics Letters*, *163*, 46-49.

Fischbacher, U. & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525-547.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42.

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), 496.

Garbarino, E., Slonim, R., & Villeval, M. C. (2019). Loss aversion and lying behavior. *Journal of Economic Behavior & Organization*, *158*, 379-393.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384-394

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114-125.

Houser, D., Vetter, S. & Winter, J. (2012). Fairness and cheating. *European Economic Review*, *56*(8), 1645-1655.

Hugh-Jones, D. (2016). Honesty, beliefs about honesty, and economic growth in 15 countries. *Journal of Economic Behavior & Organization*, *127*, 99-114.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*(1), 1-9.

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2019). Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science*, *14*(5), 778-796.

Muehlheusser, G., Roider, A., & Wallmeier, N. (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters*, *128*, 25-29.

Potters, J., & Stoop, J. (2016). Do cheaters in the lab also cheat in the field? *European Economic Review*, *87*, 26-33.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453-469.

Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, *45*, 181-196.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, *117*(523), 1243-1259.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological science*, *23*(10), 1264-1270.

FIGURE 1: # Heads Reported (Panel A) and Response Time (Panel B) cumulative distribution functions

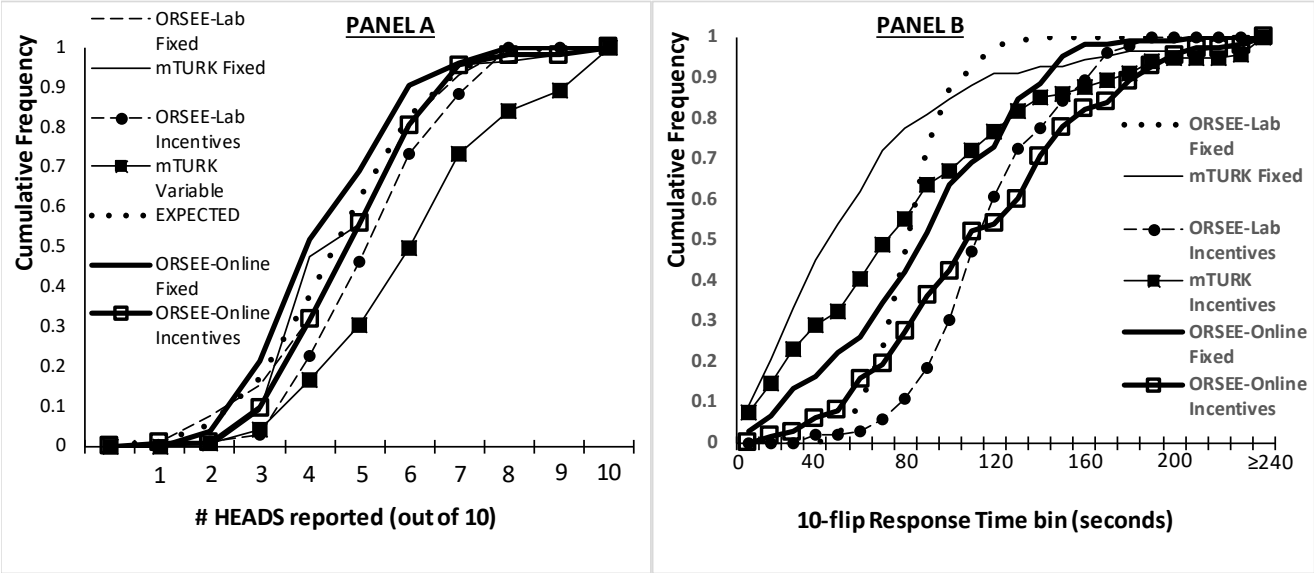
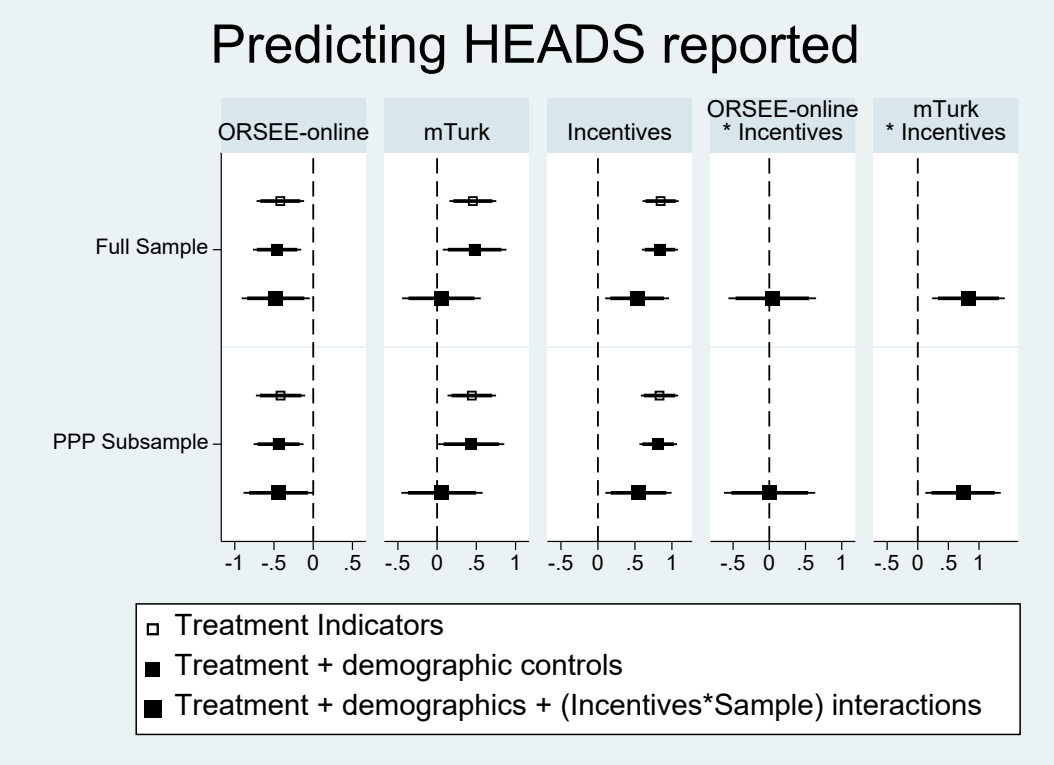
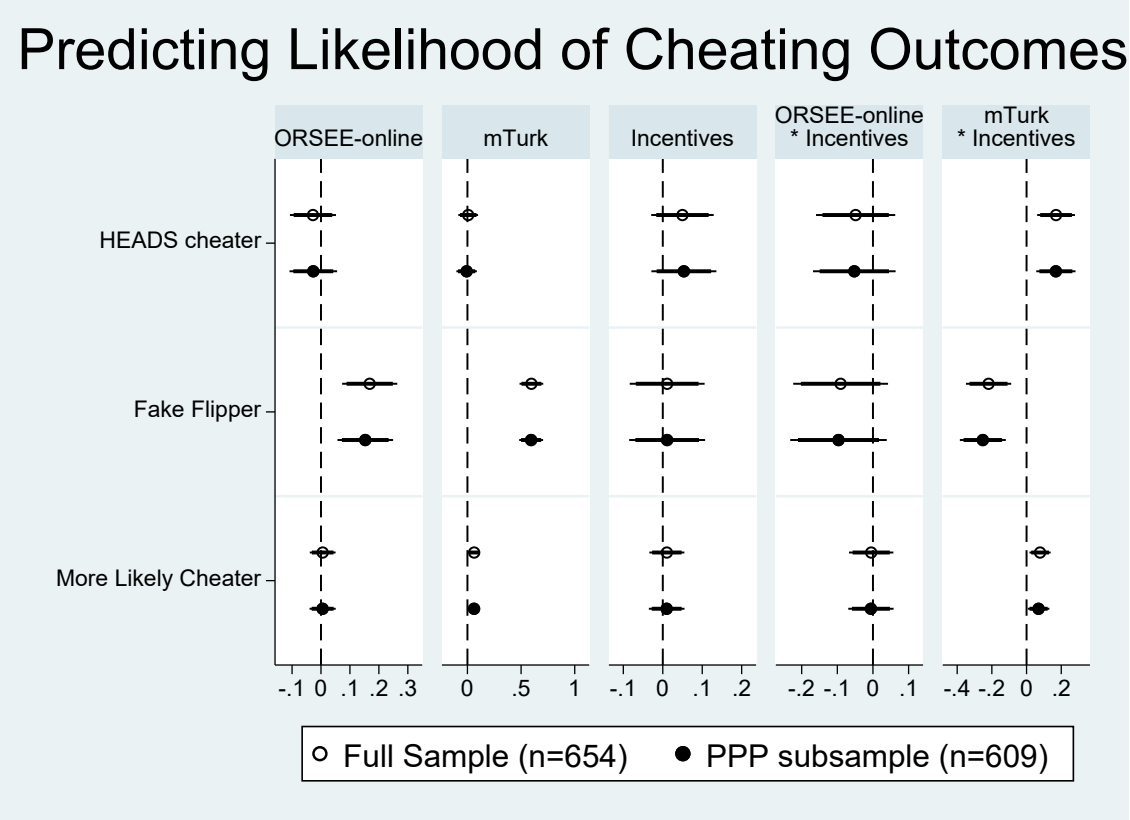


FIGURE 2: Coefficient plot of HEADS predictions



Note: See Appendix Table A1 for full estimation results. Coefficient plots show point estimates for key indicators (each column) from 3 specifications estimated on both the Full and PPP subsamples of data. Point estimates include the 90% (thicker line) and 95% (thinner line) confidence intervals around each point estimate.

FIGURE 3: Coefficient plot of Cheater likelihood predictions



Note: See Appendix Table A2 for full estimation results. Note different scales to x-axis across coefficient estimates. Coefficient plots show point estimates for key indicators (each column) from modeling the 3 categories of possible cheaters for both the Full and PPP subsamples of data. All models included preregistered demographic controls (age, gender, minority, cognitive reflection score). Point estimates include the 90% (thicker line) and 95% (thinner line) confidence intervals around each point estimate.

Table 1: Predicting Response Times

Dep Variable = Response Time (sec)	**Full Sample (n=654)**			**Passed Poison Pill (n=609)**		
Variable	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)
Constant	83.836 (4.155)***	67.521 (8.488)***	66.199 (9.000)***	83.674 (4.310)***	69.043 (8.740)***	68.234 (99.287)***
ORSEE-online	2.022 (5.062)	.256 (5.108)	2.585 (7.259)	2.936 (5.269)	1.805 (5.307)	4.623 (7.497)
mTurk	-24.823 (4.982)***	-29.888 (6.746)***	-28.250 (8.400)***	-21.968 (5.179)***	-26.490 (7.018)***	-26.748 (8.695)***
Incentives	29.676 (4.066)***	29.569 (4.062)***	32.270 (7.235)***	29.791 (4.232)***	29.735 (4.240)***	31.428 (7.508)***
ORSEE-online*Incentives	---	---	-4.594 (10.143	---	---	-5.610 (10.566)
mTurk*Incentives	---	---	-3.356 (9.940)	---	---	.336 (10.322)
Female	---	6.585 (4.338)	6.468 (4.353)	---	5.161 (4.537)	4.995 (4.551)
Minority	---	8.889 (4.952)*	8.875 (4.959)*	---	13.817 (5.256)***	13.853 (5.263)***
Age	---	.215 (.315)	.215 (.315)	---	.214 (.323)	.209 (.324)
CRT-score	---	2.469 (1.042)**	2.489 (1.047)**	---	1.690 (1.092)	1.759 (1.099)
R-squared	.121	.135	.135	.114	.128	.129

Notes: *$p < .10$, ** $p < .05$, ***$p < .01$ for the 2-tailed tests except for test of key hypotheses. P-values reported for key hypothesis regarding the effects of *ORSEE-online, mTurk*, *Incentives* and the *Incentives* interaction terms are 1-tailed p-values, as was indicated in the preregistered analysis plan. The PPP subsample also omits n=6 participants who were recruited to the *ORSEE-online Fixed* treatment but were inadvertently given the survey link to the *ORSEE-online Variable* treatment.

TABLE 2: Predicting the Likelihood of the Reported Coin Flip Sequence

Dep Variable = Sequence Probabilty (type 1 error of rejecting H_0: fair coin sequence)	**Full Sample (n=637)** (omits those reporting 10 Heads)			**Passed Poison Pill (n=592)** (omits those reporting 10 Heads)		
Variable	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)
Constant	.302 (.012)***	.264 (.024)***	.261 (.026)***	.297 (.012)***	.262 (.025)***	.265 (.026)***
ORSEE-online	-.0008 (.0143)	-.005 (.015)	-.003 (.021)	.002 (.015)	-.002 (.015)	-.006 (.021)
mTurk	-.037 (.0143)***	-.048 (.020)***	-.042 (.024)**	-.023 (.015)*	-.031 (.020)*	-.034 (.025)*
Incentives	-.005 (.012)	-.005 (.012)	.0006 (.021)	.003 (.012)	.004 (.012)	-.001 (.021)
ORSEE-online * Incentives	---	---	-.005 (.029)	---	---	.009 (.039)
mTurk*Incentives	---	---	-.011 (.029)	---	---	.006 (.029)
Female	---	.020 (.012)	.020 (.012)	---	.019 (.013)	.019 (.013)
Minority	---	.018 (.014)	.018 (.014)	---	.026 (.015)*	.026 (.015)*
Age	---	.0004 (.0009)	.0004 (.0009)	---	.0004 (.0009)	.0004 (.0009)
CRT-score	---	.005 (.003)*	.005 (.003)*	---	.004 (.003)	.004 (.003)
R-squared	.014	.025	.025	.006	.017	.017

Notes: *$p < .10$, ** $p < .05$, ***$p < .01$ for the 2-tailed tests except for test of key hypotheses. P-values reported for key hypothesis regarding the effects of *ORSEE-online, mTurk*, *Incentives* and the *Incentives* interaction terms are 1-tailed p-values, as was indicated in the preregistered analysis plan. Data from participants reporting 10 HEADS are omitted because the result in only one possible sequence (and so the runs test is not helpful in these instances). The PPP subsample also omits n=6 participants who were recruited to the *ORSEE-online Fixed* treatment but were inadvertently given the survey link to the *ORSEE-online Variable* treatment.

APPENDIX A: Full estimation results for in-text coefficient plots

TABLE A1: Predicting # Heads Reported (see Figure 2 in text)

Dep Variable = # Reported Heads	**Full Sample (n=654)**			**Passed Poison Pill (n=609)**		
Variable	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)
Constant	4.969 (.127)***	4.992 (.260)***	5.14 (.273)***	5.005 (.131)***	5.044 (.270)***	5.182 (.282)***
ORSEE-online	-.419 (.154)***	-.461 (.156)***	-.478 (.221)**	-.416 (.160)***	-.441 (.162)***	-.441 (.228)*
mTurk	.456 (.152)***	.478 (.207)**	.056 (.255)	.442 (.157)***	.435 (.214)**	.062 (.264)
Incentives	.848 (.125)***	.839 (.124)***	.530 (.220)***	.833 (.129)***	.815 (.129)***	.548 (.228)***
ORSEE-online*Incentives	---	---	.040 (.308)	---	---	.005 (.321)
mTurk*Incentives	---	---	.828 (.302)***	---	---	.740 (.313)***
Female	---	.052 (.133)	.047 (.132)	---	-.025 (.161)	-.032 (.138)
Minority	---	-.258 (.152)*	-.253 (.151)	---	-.215 (.161)	-.209 (.160)
Age	---	-.005 (.010)	-.006 (.010)	---	-.003 (.010)	-.004 (.010)
CRT-score	---	.046 (.032)	.052 (.032)	---	.038 (.033)	.043 (.033)
R-squared	.111	.118	.132	.107	.113	.124

Notes: *$p < .10$, ** $p < .05$, ***$p < .01$ for the 2-tailed tests except for test of key hypotheses. P-values reported for key hypothesis regarding the effects of *ORSEE-online, mTurk*, *Incentives* and the *Incentives* interaction terms are 1-tailed p-values, as was indicated in the preregistered analysis plan. The PPP subsample also omits n=6 participants who were recruited to the *ORSEE-online Fixed* treatment but were inadvertently given the survey link to the *ORSEE-online Incentives* treatment.

TABLE A2: Linear Probability models—predicting the likelihood of identification as a cheater
(full estimation results for coefficient plots in Figure 3 of main text)

Dep Variable = Dichotomous Indicator	**DV=*HEADS cheater***		**DV=*Fake Flipper***		**DV=*More Likely Cheater***	
	Full Sample (n=654)	PPP (n-609)	Full Sample (n=654)	PPP (n-609)	Full Sample (n=654)	PPP (n-609)
Variable	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)	Coef. (std. err.)
Constant	.131 (.050)***	.141 (.052)***	.182 (.060)***	.173 (.061)***	.086 (.028)***	.083 (.028)***
ORSEE-online	-.028 (.040)	-.027 (.042)	.168 (.048)***	.153 (.049)***	.006 (.023)	.006 (.023)
mTurk	.007 (.047)	-.007 (.049)	.596 (.056)***	.594 (.057)***	.063 (.026)***	.062 (.027)**
Incentives	.050 (.040)	.053 (.042)	.011 (.048)	.011 (.049)	.010 (.023)	.010 (.023)
ORSEE-online * Incentives	-.049 (.057)	-.052 (.059)	-.091 (.068)	-.097 (.069)	-.005 (.032)	-.006 (.032)
mTurk*Incentives	.170 (.055)***	.169 (.058)***	-.219 (.067)***	-.252 (.067)***	.078 (.031)***	.069 (.032)**
Female	-.013 (.020)	-.028 (.025)	-.006 (.029)	.003 (.030)	-.029 (.014)**	-.031 (.014)**
Minority	-.031 (.028)	-.022 (.029)	-.031 (.033)	-.057 (.034)*	-.029 (.015)*	-.021 (.016)
Age	-.002 (.002)	-.002 (.002)	-.005 (.002)**	-.006 (.002)***	-.003 (.001)***	-.003 (.001)***
CRT-score	.001 (.006)	-.0001 (.006)	-.021 (.007)***	-.012 (.007)	-.004 (.003)	-.002 (.003)
R-squared	.081	.079	.223	.223	.086	.079

Notes: *$p < .10$, ** $p < .05$, ***$p < .01$ for the 2-tailed tests except for test of key hypotheses. P-values reported for key hypothesis regarding the effects of *ORSEE-online, mTurk*, *Incentives* and the *Incentives* interaction terms are 1-tailed p-values, as was indicated in the preregistered analysis plan. Data from participants reporting 10 HEADS are omitted because the result in only one possible sequence (and so the runs test is not helpful in these instances). The PPP subsample also omits n=6 participants who were recruited to the *ORSEE-online Fixed* treatment but were inadvertently given the survey link to the *ORSEE-online Variable* treatment. Qualitatively similar results are found using nonlinear Probit estimations (results available on request).

Appendix B: Experimental Instructions

Note: All participants, *ORSEE-Lab, ORSEE-online,* and *mTurk* completed the task using this same online survey approach. That is, even the *ORSEE-Lab* treatment participants who were physically in our lab used this same survey platform to input all data and coin flip responses. Instructions below show differentiated text for those in the *Fixed* versus *Incentives* payment treatments.

Survey page 1: Consent page (tailored depending on *mTurk*, *ORSEE-online*, or *ORSEE-Lab* participants. (*ORSEE-online*** participants were then asked to input a valid email for receipt of Amazon gift code within 48 hours of survey completion).

--What is your Gender? (Male, Female)

--What is your ethnicity? ("Hispanic or Latino", or "Not Hispanic or Latino")

--What is your racial category? (American Indian/Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Black or African American, White (Caucasian), or Mixed)

--What is your age? (slider bar response in 18-100 years of age range)

----------page break ----------

--Poison Pill Question (for PPP subsample determination)

At times, individuals do these surveys without really paying attention (i.e., they go through fast just to get paid). In order to check that you are, in fact, reading through the questions, we ask that you place the slider below on the sum of thirteen and twenty-two (i.e., add those two numbers up, and what you get is where you should place the slider.....be careful that the slider does not move slightly when you make your selection and release your mouse click). If we cannot document that you are paying attention to the questions, then you may not get paid for participation.

(slider bar response along number bar from 0-100. Required response of 35 to be in PPP subsample)

----------page break ----------

--Coin Flip Task (highlights were present in survey text)

This next question asks you to flip a coin 10 times and report the results (making note of the order of Heads and Tails outcomes). Your payment in this experiment does **not** depend on the outcomes of the coin flips. **Please read the instructions on the next page carefully**. **Before advancing to the next page, first please get/find a quarter that you can flip and something to write with, and then click the button below (one should be available at your computer station). Please do not flip the coin before advancing to the next page.**

☐ I have read the instructions above, I have a quarter/coin and something to write with, and I am ready to start flipping once I click this button

----------page break ----------

(this page is timed to generate response time data)

This is a description of your compensation in this study, so please read carefully.

[FIXED treatment]

Please take your coin and **Flip the coin 10 times** and record what you get for each flip along with the total number of HEADS you flip. Please only perform the 10 coin flips once. **Your payoff for this task does not depend on the outcome of your coin flips, but we want to know the outcome of your coin flips**. For example, if you flip HEADS 3 times, your compensation will be the same as if you flip HEADS 6 times (or any other number of times).

Please tell us below how many times you flipped HEADS out of your 10 total coin flips.

[INCENTIVES treatment]

Please take your coin and **Flip the coin 10 times** and record what you get for each flip along with the total number of HEADS you flip. Please only perform the 10 coin flips once. **Your payoff for this task will be an additional $0.25 (i.e., 25 cents) for each HEADS outcome from these coin flips**, and this payment will be added to your $5.00 fixed Amazon gift code compensation for this experiment. For example, if you flip HEADS 3 times, your additional compensation (on top of the $5.00 fixed payment) will be another $0.75. Similarly, if you flip HEADS 6 times you will receive an additional $1.50, etc.

Please tell us below how many times you flipped HEADS out of your 10 total coin flips, and we will add 25 cents for each HEADs flipped to your fixed payment for today's experiment.

(Slider bar response to input number of HEADS flipped from 0 to 10)

----------page break ----------

Please report the outcomes for each of the 10 coin flips in the order in which they occurred.

(10 rows of multiple choice options between HEADS and TAILS to indicate each flip by number)

----------page break ----------

Finally, please answer the questions on the next set of pages for us.

----------page break ----------

6-Item Cognitive Reflection Task (Primi et al, 2016)

Question: A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (please indicate your numeric answer **in cents**. For example, 30 cents would be "30", not ".30", 1 cents would be "1" and not ".01", etc)……correct = 5 cents

----------page break ----------

Question: It takes 5 minutes for 5 machines to make 5 widgets, how long would it take for 100 machines to make 100 widgets? (please indicate your numeric answer **in minutes**)…..correct = 5 minutes

----------page break ----------

Question: In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover **half** the lake? (please indicate your numeric answer **in days**)…..correct = 47 days

----------page break ----------

Question: If 3 elves can wrap 3 toys in 1 hour, how many elves are needed to wrap 6 toys in 2 hours? (please give your numeric answer in **# of elves**)…..correct = 3 elves

----------page break ----------

Question: Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are there in the class? (please give your numeric answer in **# of students**)….correct = 29 students

----------page break ----------

Question: In an athletics team, tall members are **three** times more likely to win a medal than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes? (please give your numeric answer in **# of medals**)……correct = 15 medals