Journal of Experimental Political Science **5** (2018) 217–229
doi:10.1017/XPS.2018.7

Pay Rates and Subject Performance in Social Science Experiments Using Crowdsourced Online Samples*

David J. Andersen* **and Richard R. Lau**†

Abstract

Mechanical Turk has become an important source of subjects for social science experiments, providing a low-cost alternative to the convenience of using undergraduates while avoiding the expense of drawing fully representative samples. However, we know little about how the rates we pay to "Turkers" for participating in social science experiments affects their participation. This study examines subject performance using two experiments – a short survey experiment and a longer dynamic process tracing study of political campaigns – that recruited Turkers at different rates of pay. Looking at demographics and using measures of attention, engagement and evaluation of the candidates, we find no effects of pay rates upon subject recruitment or participation. We conclude by discussing implications and ethical standards of pay.

Keywords: Mechanical Turk, experimental design, crowdsourced samples, pay rates

"Crowdsourcing" samples have emerged as a fast, easy, and inexpensive source of subjects for experimental research. In particular, Amazon's Mechanical Turk has become a popular source for quickly and cheaply recruiting large numbers of respondents (Berinsky et al., 2012; Paolacci et al., 2010). "Turkers," as they are known, are a ready alternative to undergraduates or professionally assembled samples, and offer two major benefits: their availability (Hitlin, 2016; though also see Stewart et al., 2015) and their inexpensive cost, while still providing a diverse pool of subjects (Huff and Tingley, 2015; Ipeirotis, 2010; Levay et al., 2016).

The data, code, and any additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at: doi:10.7910/DVN/VCWWGZ.

*Department of Political Science, Iowa State University, 547 Ross Hall, Ames, IA 50011-1204, USA, e-mail: dander@iastate.edu
†Department of Political Science, Rutgers University, 89 George Street, New Brunswick, NJ 08901-1411, USA, e-mail: ricklau@polisci.rutgers.edu

Determining what to pay subjects on Mechanical Turk can be challenging for two reasons that may risk the quality of the sample recruited. First, different pay rates may attract different participants. Turkers selectively choose which available HITs they will accept, making it possible that the selection process may introduce sample biases (Krupnikov and Levine, 2014). Higher pay rates may attract a different type of worker than lower pay rates, either demographically or along some other factor that might influence subject performance. Second, paying too little in compensation may lead to sub-par subject attention, as participants who decide they are not going to be sufficiently compensated alter their performance (Berinsky et al., 2016).

For simple tasks with "right" or "wrong" results that the Requester can evaluate, there is an easy mechanism for evaluating subject behavior – rewarding accurate behavior through payment and punishing inaccurate behavior by denying payment. The Requester simply checks on the work as it is returned to make sure that the Worker was indeed paying attention and performing adequately. Turkers know this, and behave accordingly.

As Ho, Slivkins, Suri, and Vaughan describe: "even when standard, unconditional payments are used and no explicit acceptance criteria is specified, workers may behave as if the payments are *implicitly* performance-based since they believe their work may be rejected if its quality is sufficiently low" (Ho et al., 2015). In such scenarios, different pay rates have been demonstrated to motivate workers to do a greater quantity of work, but not at higher quality (Mason and Watts, 2009). Similarly, several studies have shown that, when work is verifiable based upon accuracy or correctness, pay rates can influence worker behavior positively (Finnerty et al., 2013; Horton and Chilton, 2010; Ho et al., 2015; Ye et al., 2017).

Social scientists should take pause at this, because all of these studies are conditional upon the ability to review subject performance using objective criteria. For example, determining if a subject correctly ordered images, or successfully identified words among a jumble of letters is relatively easy (Mason and Watts, 2009). However, subject performance in social scientific studies tends to lack a strong evaluation component. That is, subjects are asked to behave "normally" and react to the information and stimuli they are provided as they would in the real-world, but without the ability of the experimenter to verify that they are indeed doing so. Behaving "normally" does not clearly indicate a "right" or "wrong" set of behaviors that can be observed. It is exceedingly difficult to determine if a subject is paying attention to an online study (Berinsky et al., 2012; Berinsky et al., 2016; Hauser and Schwarz, 2016, Paolacci et al., 2010), or answering honestly (Chandler et al., 2014; Rouse, 2015) or behaving as they normally would.

METHOD

We identified three areas where payment might affect subject behavior that could matter to a researcher: self-selection (who chooses to accept the HIT), engagement

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

(how actively subjects paid attention to and interacted with the study), and performance (how those subjects reacted to what they saw in the study). Since, we can identify no correct form of behavior; we simply look to see if different pay rates produce *different* between-subject behavior across a range of measures. If pay rates do play an influence, we would expect to see either a linear relationship (where higher rates of pay lead to greater attention and performance), or a threshold effect (where performance shifts when an "acceptable rate" has been reached) on a consistent basis. Thus, we are not seeking a single significant finding, but are looking for emerging patterns of behavioral differences that emerge between pay groups.

We conducted two separate studies – one short and easy, the other long and difficult – in order to view the effects of different pay rates on performance in different styles of social science experiments. The first study was a short survey experiment designed in Qualtrics, involving one randomized image followed by 13 questions.[1] The second study was programmed in the Dynamic Process Tracing Environment (DPTE) and asked subjects to learn about and vote for political candidates.[2]

If pay rates influence subject recruitment and participation, we anticipate subjects are likely to perform optimally when their compensation is highest (Hus et al., 2017; Ye et al., 2017). Subjects who feel they are being adequately compensated for their work are more likely to pay attention, to take seriously the task at hand, and to focus on the decisions they are asked to consider. Of course, as the studies progress and subjects spent greater time and effort in participating, their attitudes about "being adequately compensated" may change.

Thus, we further suspect that any differences in subject behavior are more likely to show up later in the study than earlier. Our first study, which took only about 4 min to complete, was unlikely to produce differences in behavior between the beginning and end of the survey. Our second study however, which could take 60 minutes to complete, we believe is more likely to produce effects toward the end of the study as subjects tired of participation and may have begun re-evaluating whether their payment was indeed adequate.

RESULTS

Our results from both studies were roughly identical, in that we found few reportable differences in our measures between the different pay rates.[3] For brevity,

[1] The study can be viewed at: https://iastate.qualtrics.com/jfe/form/SV_1YxsPYdlywrENi5

[2] A more thorough description of the study can be found in the online appendix. The HIT we posted and the full study we employed can be viewed online at: https://dpte.polisci.uiowa.edu/dpte/action/player/launch/921/22772?pass=Archived&skip=1

[3] The data, syntax, and additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at: doi:10.7910/DVN/VCWWGZ

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

and to save space on reproducing dozens of null results, we only present our second study here, as it permits the more thorough look at Turker behavior. Matching results for the survey experiment can be found in the Online Appendix.

We first examine if our pay rates affected who we recruited to complete our study. We had no *a priori* assumptions about how pay rates might affect recruitment, so we relied on what we considered to be "conventional" demographic measures that we use in political science.

Table 1[4] shows that none of our eight categories (percentages of women, African-Americans, Hispanics, Democrats, Independents, or the mean age, political interest, or conservatism of our subjects) return significant results. Further, only one of our categories shows a consistent pattern in the results (a steady increase in Hispanic subjects as pay rates increased). With a relatively small sample size of 364 subjects, it is possible that a larger sample size might produce significant results, but looking at the substantive differences in results, it seems more likely that our demographic measures tended to show random fluctuation between the pay rates, rather than systematic differences in who chose to sign up for the study.

Our larger concern is for things that we were not able to measure, such as Turker experience. It is possible that more experienced Turkers may gravitate toward higher pay rates, or studies that they feel have a higher pay-to-effort ratio. This is, regrettably, something that we were not able to measure. However, since experimental samples do not tend to seek representative samples on Mechanical Turk, we feel that the risk of any demographic or background differences in who we recruit is that it could then lead to differences in behavior, either through attention to the study or in reaction to the various elements of the study. While we do not find observable demographic differences, we can continue on by examining how people performed within the study.

An advantage of using a DPTE experiment is that we have much greater ability to tease out how subjects performed across a range of measures. We first present the results of our attention checks, and then will move on to discuss engagement with the experiment and candidate evaluation.

Table 2 shows that the vast majority of all of our subjects passed our attention check tests, and there are again no significant differences between our pay rate groups.[5] There is an apparent pattern of subjects passing at higher rates when paid more however, which suggests that perhaps there may be an effect that our study

[4] Pay rates could also influence how fast subjects accept and complete the study, but we found no evidence of this. Every batch we posted completed in approximately the same time, but because of the nature of how AMT posts HITs and reports completions, it is difficult to analyze more precisely. The lower pay rate groups closed slighty slower than the higher pay rate groups, but the substantive difference was minimal and seemed to be caused by subjects accepting the HIT and then waiting to complete it until the time limit was due.

[5] Due to a programming glitch, our subjects on the $2 pay day did not see the attention check questions, but they did still view our "pop up" attention checks.

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

Table 1
Subject Demographics of the DPTE study, by Pay Rate

Pay Rate	% Female	% Black	% Hispanic	% Democrat	% Indepen.	Mean Age	Mean Pol. Int.	Mean LibCon
\$2 (*n*=99)	53.6%	4.0%	8.2%	61.2%	15.3%	35.22 (1.23)	2.14 (0.07)	3.44 (0.16)
\$4 (*n*=96)	46.8%	6.3%	9.6%	68.1%	9.6%	33.88 (1.11)	2.19 (0.07)	3.16 (0.18)
\$6 (*n*=99)	35.4%	9.1%	14.1%	56.6%	14.1%	31.87 (0.97)	2.17 (0.07)	3.33 (0.17)
\$8 (*n*=70)	49.3%	5.7%	14.5%	63.8%	17.4%	32.20 (1.17)	1.99 (0.10)	3.53 (0.19)
Total (*n*=364)	46.0%	6.3%	11.4%	62.2%	13.9%	33.37 (0.57)	2.13 (0.04)	3.35 (0.09)
Pearson χ^2	7.101	2.197	2.719	2.834	2.341			
F Statistic						1.975	1.253	0.775

Table 2
Subject Reaction to Attention Checks, by Pay Rate

		Primary Election				General Election			
Pay Rate	Pass Trap Qs	Pass PopUp 1	Pass PopUp 2	Pass PopUp 3	Pass PopUp 4	Pass PopUp 1	Pass PopUp 2	Pass PopUp 3	Pass PopUp 4
\$2 (*n*=99)	–	93.8%	93.8%	88.2%	85.7%	94.8%	100.0%	94.4%	97.8%
\$4 (*n*=96)	93.6%	96.8%	98.9%	100.0%	100.0%	97.9%	93.8%	96.2%	93.6%
\$6 (*n*=99)	94.9%	94.9%	96.0%	100.0%	100.0%	94.9%	100.0%	100.0%	92.5%
\$8 (*n*=70)	91.2%	95.7%	92.8%	100.0%	100.0%	98.6%	100.0%	97.2%	100.0%
Total (*n*=364)	93.5%	95.3%	95.5%	96.0%	94.7%	96.4%	97.9%	97.0%	95.8%
Pearson χ^2	0.947	0.998	4.523	4.044	1.810	2.766	2.043	3.140	3.606

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

was not large enough to fully capture. The lowest rates of passing the first two popups in the Primary are found in the $2 pay group (93.8% for both), and while subjects in the higher pay groups all passed the third and fourth popup at a 100% rate, subjects in our minimal $2 pay group passed this at the lowest rates we find in the study, below 90%. While not a significant finding, this suggests that perhaps subjects in this lowest pay group were not paying attention to the extent of the other pay groups.

If this is the case, however, further evidence should emerge elsewhere. We would expect that attention would get worse as the study carried on. However, it does not. These differences do not appear again in the General Election, when we expected effects to be the greatest. Overall, we find that our subjects generally responded well to our attention checks regardless of what they were being paid.

Beyond merely paying attention to what was presented to them, this study also asked subjects to actively engage with the program, and actively learn about political candidates. This is another area where differential motivation based upon pay rates could influence behavior. Table 3 presents a series of one-way analysis-of-variance tests on measures of active engagement with the experiment. While the previous table measured how much attention subjects paid to the study, this table assesses how actively engaged Turkers were in interacting with the dynamic information boards by selecting information to view. If payments created different incentives to participate, this should be observable through the time subjects spent in the campaign scenarios, the number of items they chose to view, and how much time they devoted to the political aspects of the study relative to the more entertaining current event items.

We find only one statistically significant result, and thus no consistent or clear evidence that pay rates influenced our subject behavior. The lone significant finding we have occurs for our measure of the number of information items subjects chose to open during the Primary Election. While significant, these results show that our highest paid group sought out the most information in the primary, while the second highest group sought out the least. This does not sensibly fit to our theory, and is not replicated along other measures. The lack of a clear pattern within the data again suggests that pay rates did not systematically influence subject performance, even in a long and taxing study.

A final way for us to consider how our subjects participated in the study is to evaluate their final decisions and evaluations of the candidates. It is possible that, while behavioral differences did not emerge, perhaps psychological appraisals of the subject matter were effected by anticipated rewards. We find, again, very little evidence that pay rates mattered. We asked our subjects who they voted for, how confident they were in their vote decision, how difficult that vote choice was, and how much they felt they knew about the candidates, for both the Executive and House race.

The only significant finding we have in Table 4 is for the confidence our subjects had in selecting the House candidate that they truly preferred. Here, we find a

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

Table 3
Subject Engagement with the Experiment, by Pay Rate

	Primary Election					General Election				
Pay Rate	Total Time	Avg # of Items Viewed	Avg Time Viewing Items	Avg Time Viewing Pol Items	Avg Time Viewing CE Items	Total Time	Avg # of Items Viewed	Avg Time Viewing Items	Avg Time Viewing Pol Items	Avg Time Viewing CE Items
$2 (*n*=99)	530.07 (20.16)	35.45 (1.84)	228.69 (17.33)	197.89 (10.00)	30.80 (10.59)	428.30 (17.94)	34.58 (2.05)	196.34 (11.06)	183.67 (10.84)	12.67 (2.46)
$4 (*n*=96)	477.78 (12.94)	35.00 (2.06)	226.43 (12.88)	207.78 (12.56)	18.65 (2.60)	393.61 (10.54)	33.13 (1.96)	206.09 (12.01)	194.88 (12.05)	11.21 (1.74)
$6 (*n*=99)	474.64 (12.45)	31.78 (1.72)	203.66 (10.14)	181.48 (9.27)	22. 18 (3.41)	389.18 (9.61)	31.80 (1.75)	183.64 (8.96)	168.88 (8.58)	14.76 (2.08)
$8 (*n*=70)	486.71 (20.24)	42.77 (4.68)	212.35 (14.29)	187.91 (13.61)	24.44 (3.97)	382.41 (11.74)	36.22 (2.87)	184.19 (12.81)	168.84 (12.87)	15.36 (2.69)
Total (*n*=364)	493.01 (8.36)	35.73 (1.26)	218.01 (6.98)	193.96 (5.60)	24.05 (3.19)	399.71 (6.65)	33.75 (1.05)	192.98 (5.55)	179.59 (5.50)	13.40 (1.12)
F Stat Sig	2.603	2.973*	0.770	1.101	0.681	2.461	0.754	0.932	1.314	0.701

Table 4
Subject Evaluation of the Candidates, by Pay Rate

Pay Rate	Exec Dem Vote	Exec Vote Conf	Exec Vote Diff	Exec Cand Know	Hse Dem Vote	House Vote Conf	House Vote Diff	Hse Cand Know	Avg Cand Pref
\$2 (*n*=99)	66.0%	3.804 (0.109)	2.289 (0.129)	2.938 (0.073)	63.9%	3.897 (0.113)	2.289 (0.139)	2.691 (0.088)	33.26 (2.14)
\$4 (*n*=96)	66.7%	3.776 (0.114)	2.277 (0.126)	2.920 (0.069)	67.7%	3.702 (0.119)	2.351 (0.126)	2.700 (0.079)	29.63 (2.16)
\$6 (*n*=99)	62.6%	3.816 (0.115)	2.010 (0.107)	3.040 (0.075)	59.6%	3.612 (0.104)	2.141 (0.098)	2.722 (0.087)	29.39 (1.96)
\$8 (*n*=70)	68.1%	3.427 (0.138)	2.318 (0.150)	2.862 (0.088)	65.2%	3.368 (0.145)	2.603 (0.144)	2.486 (0.101)	25.46 (2.27)
Total (*n*=364)	65.6%	3.728 (0.059)	2.214 (0.063)	2.947 (0.038)	64.0%	3.667 (0.060)	2.324 (0.062)	2.662 (0.044)	29.74 (1.07)
Pearson Chi^2	0.635				1.443				
F Statistic		2.073	1.345	0.932		3.098*	2.151	1.299	2.030

significant result and a pattern indicating that lower-paid subjects had greater confidence in their vote choice. This could lead us to assume that our rates of pay influenced how much consideration or psychological investment our subjects had in the study. However, this again appears to be an isolated finding. In all other measures, there are no significant differences or patterns in the data to find that pay rates played a role in how our subjects felt about the candidates or their vote decisions.

CONCLUSIONS

Our results are quite easy to summarize – pay rates did not seem to matter much to subject performance among Mechanical Turkers, at least not that we observed. While we only discuss our first study here, these results are replicated across another shorter study that collected a much larger sample and is presented in the Online Appendix. In both studies, no systematic patterns emerged that might suggest that pay rates significantly or substantively influenced subject behavior. This does not mean, of course, that pay rates produce no effects, but simply that we, using two very different social science studies, and observing numerous measures of behavior in each, were not able to identify any such effects. We do feel that have observed most, if not all, of the important characteristics of behavior likely to change.

Importantly, we report these results without correcting for multiple hypotheses testing, which would only further reduce the minimal effects we found. In each of our four areas, we analyze we have at least eight different measures, suggesting that by chance alone we should find some significant findings. Indeed, we do. However, these findings show no clear patterns of the influence of pay rates and it is in the absence of patterns that we feel safest in drawing our conclusions. Our clearest path is to conclude that pay rates largely do not influence subject participation and behavior on Mechanical Turk.

This is an important null finding for social scientists using online labor pools. However, we do not intend here to conclude fully that pay rates do not matter. Paying a fair wage for work done does still involve ethical standards (Zechmeister, 2015). While our discipline as a whole has never established what ethical wages are for subjects, several suggestions both within the Turker community and academic literature have suggested a $6 per hour rate. This still makes crowdsourced samples considerably cheaper than professional alternatives, while also paying a fair rate to the people whose work we depend upon.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/XPS.2018.7

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

REFERENCES

Andersen, David. 2018. "Replication Data for: Subject Performance in Social Science Experiments Using Crowdsources Online Samples." doi:10.7910/DVN/VCWWGZ, Harvard Dataverse, V1, UNF:6:RQAq0OAZinHNkPjUZVcz5A==.

Berinsky, Adam, Gregory Huber, and Gabriel Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20: 351–368.

Berinsky, Adam, Michele Margolis, and Michael Sances. 2016. "Can we turn shirkers into workers?." *Journal of Experimental Social Psychology* 66: 20–28.

Druckman, James N. and Cindy D. Kam. 2011. "Students As Experimental Participants: A Defense of the 'Narrow Data Base'." In *Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H Kuklinski, and Arthur Lupia. (pp. 41–57). New York: Cambridge University Press.

Druckman, James, Donald Green James Kuklinski, and Arthur Lupia 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4): 627–635.

Finnerty, Ailbhe, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. "Keep it Simple: Reward and Task Design in Crowdsourcing." Paper presented at CHItaly '13, Trento, Italy, September 16–20.

Hauser, David J. and Norbert Schwarz. 2016. "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks Than do Subject Pool Participants." *Bahavior Research Methods* 48(1): 400–407.

Hitlin, Paul. 2016. 'Research in the Crowdsourcing Age, a Case Study' Pew Research Center. July 2016. Available at: http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

Ho, Chien-Ju, Aleksandrs Slivkins, Diddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. Paper presented at the International World Wide Web Conference, Florence, Italy, May 18–22.

Horton, John J. and Lydia B. Chilton. 2010. "The Labor Economics of Paid Crowdsourcing." Presented at the 11th ACM conference on electronic commerce (pp. 209–218). Cambridge, Massachusetts: ACM.

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).

Hus, Joanne W., Maximilian D. Schmeiser, Catherine Haggerty, and Shannon Nelson. 2017. "The Effect of Large Monetary Incentives on Survey Completion: Evidence from a Randomized Experiment with the Survey of Consumer Finances." *Public Opinion Quarterly* 81(Fall): 736–747.

Ipeirotis, Panagiotis G. 2010. Demographics of Mechanical Turk. NYU Working Paper No. CEDER-10-01. Available at SSRN: https://ssrn.com/abstract=1585030. Accessed March 14, 2018.

Iyengar, Shanto. 2011. "Laboratory Experiments in Political Science." In *Handbook of Experimental Political Science*, Eds. James Druckman, Donald Green, James Kuklinski and Arthur Lupia. New York City: Cambridge University Press.

Kaufman, Nicolas, Thimo Schulze, and Daniel Veit. 2011. "More than Fun and Money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk." Presented at the during the Proceedings of the Seventeenth Americas Conference on Information Systems. Detroit, Michigan, August 4–7.

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross-sample Comparisons and External Validity." *Journal of Experimental Political Science* 1: 59–80.

Lau, Richard R. 1995. "Information Search During an Election Campaign: Introducing a Process Tracing Methodology for Political Scientists." In *Political Judgment: Structure and Process*, Eds. M. Lodge and K. McGraw (pp. 179–206). Ann Arbor, MI: University of Michigan Press.

Lau, Richard R., David J. Andersen, and David P. Redlawsk. 2008. "An Exploration of Correct Voting in Recent Presidential Elections." *American Journal of Political Science* 52(2): 395–411.

Lau, Richard R. and David P. Redlawsk. 1997. "Voting Correctly." *American Political Science Review* 91(September): 585–599.

Lau, Richard R. and David P. Redlawsk. 2006. *How Voters Decide: Information Processing during Election Campaigns*. New York: Cambridge University Press.

Levay, Kevin E., Jeremy Freese, and Jamie Druckman. 2016. "The Demographic and Political Composition of Mechanical Turk Samples." *SAGE Open*, January-March, 2016, 1–17.

Mason, Winter and Duncan Watts. 2009. "Financial Incentives and the "Performance of Crowds." *SIGKDD Explorations* 11(2): 100–108.

McCrone, David and Frank Bechhofer. 2015. *Understanding National Identity*. Cambridge: Cambridge University Press.

McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science* 5: 31–61.

Morton, Rebecca and Kenneth Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge University Press.

Mutz, Dianna. 2011. *Population-based Survey Experiments*. Princeton, NJ: Princeton University Press.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis Ipeirotis. 2010. "Running Experiments on Mechanical Turk." *Judgment and Decision Making*, 5(5).

Rogstadius, Jakob, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets." Presented at the Fifth International AAAI Conference on Weblogs and Social Media.

Rouse, Steven V. 2015. "A Reliability Analysis of Mechanical Turk data." *Computers in Human Behavior*, 43: 304–307.

Schulze, Thimo, Simone Krug, and Martin Schader. 2012. "Workers' Task Choice in Crowdsourcing and Human Computation Markets." Presented at the thirty third International Conference on Information Systems, held in Orlando, Fl.

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences on a Narrow Data Base on Social Psychology's View on Human Nature." *Journal of Personality and Social Psychology* 51(3): 515–530.

Stewart, Neil, Cristoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newll, Gabriele Paolacci, and Jesse Chandler. 2015. "The Average Laboratory Samples a Population of 7300 Amazon Mechanical Turk Workers." *Judgement and Decision Making* 10(5): 479–491.

Ye, Teng, Sangseok You, and Lionel P. Robert 2017. "When does more Money Work? Examining the Role of Perceived Fairness in Pay on the Performance of Crowdworkers." Presented at the Eleventh International AAAI Conference on Web and Social Media.

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press

Zechmeister, Elizabeth. 2015. “Ethics and Research in Political Science: The Responsibilities of the Researcher and the Profession.” In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professional*, ed. Scott Desposato. New York, NY: Routledge.

Zizzo, Daniel. 2010. Experimenter Demand Effects in Economic Experiments. *Experimental Economics* 13(75).

https://doi.org/10.1017/XPS.2018.7 Published online by Cambridge University Press