Pay Rates and Subject Performance in Social Science Experiments Using Crowdsourced Online Samples[1]

"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 (Paolacci Et. Al. 2010; Berinsky Et. Al. 2012). "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 (Levay, Freese and Druckman, 2016; Huff and Tingley, 2015; Ipeirotis 2010).

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, Margolis and Sances 2016).

[1] 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

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. 2013). 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 (Ho et. al. 2013; Horton and Chilton, 2010; Ye, You and Robert, 2017; Finnerty et.al. 2013).

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 (Paolacci, Chandler and Ipeirotis, 2010; Berinsky, Huber and Lenz, 2012; Hauser and Schwarz, 2016, Berinsky, Margolis and Sances 2016), or answering honestly (Rouse 2015; Chandler, Mueller and Paolacci, 2014) 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 (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 thirteen questions.[2] The second study was programmed in the Dynamic Process Tracing Environment (DPTE) and asked subjects to learn about and vote for political candidates.[3]

If pay rates influence subject recruitment and participation, we anticipate subjects are likely to perform optimally when their compensation is highest (Ye, You and Robert, 2017; Hus, Schmeiser, Haggerty and Nelson 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 minutes 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 towards the end of the study as subjects tired of participation and may have begun reevaluating whether their payment was indeed adequate.

Results

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

[3] 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

Our results from both studies were roughly identical, in that we found few reportable differences in our measures between the different pay rates.[4] For brevity, 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.

<<<< Insert Table 1 here >>>>

Table 1[5] 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 show 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

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

[5] 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.

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 towards 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 don't 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 don't 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.

<<<< Insert Table 2 here >>>>

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.[6] 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 was not large

[6] 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.

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.

<<<< Insert Table 3 here >>>>

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.

<<<< Insert Table 4 here >>>>

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 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 hypothesis testing, which would only further reduce the minimal effects we found. In each of our four areas we analyze we have at least 8 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 don't matter. Paying a fair wage for work done does still involve ethical standards (Zechmeister 2013). 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.

Tables

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 Chi^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

Pay Rate	Pass Trap Qs	Primary Election				General Election			
		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 Chi^2	0.947	0.998	4.523	4.044	1.810	2.766	2.043	3.140	3.606

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

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

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, 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.

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. New York: Cambridge University Press, 41-57.

Finnerty, Ailbhe, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. Paper presented at CHItaly '13, September 16-20, 2013 in Trento, Italy.

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. ‘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, May 18-22, 2015 in Florence, Italy.

Horton, John J. and Lydia B. Chilton. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on electronic commerce*. Cambridge, Massachusetts,. ACM, 209-218.

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.

Iyengar, Shanto. 2011. Laboratory experiments in political science. In *Handbook of experimental political science.* Druckman, James, Donald Green, James Kuklinski and Arthur Lupia (Eds.) 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 during the *Proceedings of the Seventeenth Americas Conference on Information Systems.* Detroit, Michigan, August 4-7, 2011.

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 M. Lodge and K. McGraw (Eds.) *Political judgment: Structure and Process* (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 University Press. Princeton, NJ.

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. In *Proceedings of 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. In *Proceedings of the thirty third International Conference on Information Systems*, held in Orlando, Fl 2012.

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, Jr. 2017. When does more money work? Examining the role of perceived fairness in pay on the performance of crowdworkers. Accepted to the *Proceedings of the Eleventh International AAAI Conference on Web and Social Media.*

Zechmeister, Elizabeth. 2013. Ethics and Research in Political Science: The Responsibilities of the Researcher and the Profession. In Scott Desposato (Ed.) *Ethical Challenges in Political Science Experiments*.

Zizzo, Daniel. 2010. Experimenter demand effects in economic experiments. *Experimental Economics*. 13(75)

Online Appendix

Determining our pay rates

To determine what our pay rates would be, we relied upon a general convention, in both the Turker community and academic literature, that a "fair" wage on Mechanical Turk should equate to approximately $6 per hour. Setting this as our "fair hourly rate" we created payments around this of $8 (high rate of pay), $4 (low rate) and $2 (very low). We have two lower-paying groups because we believe that researchers are more likely to err on the side of underpaying samples (and concurrently increasing sample size) than overpaying. Additionally, our fears were largely of lowly-paid subjects "underperforming" by not paying attention or engaging in the study, whereas we have fewer fears of highly paid subject "overperforming" somehow. We pretested our two studies using undergraduate research assistants to determine approximate completion times, and generated pay rates based upon the estimated time to completion at the hourly rates we specified above.

In the shorter study, we found that participants ranged between 3 and 5 minutes for completion, and decided to pay subjects either $0.15, $0.30, $0.50, or $0.75, equating to roughly a rate of $2.25, $4.50, $7.50 and $11.25 per hour, for a subject who took 4 minutes to finish. The second study was longer and had much greater variance in completion time, ranging from 40 to 60 minutes, so we decided to set a rate of pay estimating the full hour for completion, and simply paid the hourly rates we generated.

Description of the two studies

The first experiment examined whether patriotic images influenced feelings of patriotism at the national and state level among Americans. It required approximately 3 minutes to complete and recruited just over 1000 Turkers. To get a nationwide sample, we recruited 4 batches of 100 subjects each simultaneously at 7:00am, 10:00am, and noon (all times Eastern), paying 15, 30, 50, and 75 cents for completion. Each batch featured the same exact description and varied only by the amount of compensation offered for completion. The rates of pay were sufficiently spaced out that, while the four pay rates were available simultaneously, they appeared on different recruitment pages.[7]

The first study was a typical survey experiment – we delivered a 13-question survey to respondents, but manipulated the image they viewed when they began. Our attempt was to provoke feelings of national patriotism, and then ask about the balance between respondents national and state-level allegiances. After viewing the image, subjects then answered a series of questions about their feelings of national patriotism, their state patriotism, and where their balance of feelings leaned. We then concluded by asking them to answer a few political knowledge questions about their state government.

[7] HIT recruitment pages are by default sorted by the rates of pay offered, putting our HITs on different screens. Mechanical Turk typically has over 1000 HITs available at any given time, ranging from pay rates of $0.01 to nearly $50, spreading these HITs out rather widely. Subjects could have sorted HITs by the time of posting which may have allowed them to see all four pay rates on the same screen, which is why we use a different method in our second study.

The second study was much longer and intensive. It took about 45-60 minutes to complete, and involved a simulated political campaign designed in the dynamic process tracing environment (DPTE). In that study we recruited a total of about 400 subjects over four consecutive days (100 per day),[8] changing the payment promised for participation each day. Again, we posted the same HIT description on Mechanical Turk, recruiting subjects with a promised payment of $2, $4, $6 or $8. We recruited subjects in batches of 25-50 at 7:00am, 10:00am, noon, and 2pm until we had 100 complete subjects at each pay rate.[9]

The second study asked subjects to learn about political candidates running for office and then vote for their preferred candidates, first in a primary election and then in a general election. We anticipated that a more difficult, cognitively-demanding study would be more likely to elicit behavioral differences from a sample responding to varying pay incentives. In other words, we figured that subjects who faced a long, difficult task may become less active, attentive or interested as the study progressed and they began to feel insufficiently compensated for their efforts. Thus, we employed a rather complex design as a tough test of the effect of pay rates, assuming that shorter, simpler studies would producer weaker effects.

[8] On our fourth day we experienced with a server error resulting in the loss of about 30 subjects' data. We see no evidence that this substantively effected our results.

[9] Assessing the effects of pay rates is complicated by the fact that the treatment (pay rates) cannot be randomly assigned to participants, because Turkers need to know the pay rate before they will accept a HIT. Instead we tried two different approaches, first by simultaneously posting 4 identical HITs at different pay rates (Study 1), and then by posting identical HITs over the course of 4 days, paying a randomizing the rate paid on each day (Study 2). Both methods have weaknesses, but together provide a strong basis to analyze the influence of differing pay rates.

The second study took between 45-60 minutes to complete, and require subjects to actively participate by learning about political candidates running for office. Subjects began the study by completing a questionnaire including approximately 100 questions, including: 35 political opinions, 11 political knowledge questions, 10 personality questions, 7 political participation questions, 24 feeling thermometer evaluations of political candidates and groups, and 9 demographic questions. Within these questions were 4 "attention check" questions – simple questions that tracked whether subjects were paying attention to what they were shown or just randomly clicking an answer (See Berinsky, Markolis and Sances 2016).

Subjects then proceeded to a short 2-minute long mock presidential campaign that taught them how the dynamic process tracing software operated. After completing that training, they were introduced to two Democrats and two Republicans running their party's nomination for an executive office,[10] and asked to choose in which primary they would vote. They then viewed a 17-minute long dynamic information board and cast their primary vote (see Lau and Redlawsk 2006 for a fuller explanation of the system). After 7 minutes, subjects were able to choose to leave the dynamic information board and proceed directly to the "ballot box" to vote.

After the primary, subjects found out which candidates won and saw those candidates advance to the general election, where they were joined by candidates

[10] Half the sample saw the candidates as gubernatorial candidates, and half saw them appear as presidential candidates. The candidates were identical in all ways other than the office they sought.

running for US Senate and the House of Representatives. Another 17-minute dynamic information board followed, after which subjects were asked to vote and evaluate the candidates they learned about. In total, subjects were able to learn up to 25 unique attributes about each candidate.

Alongside this candidate information, we included 25 current event items that were not politically relevant, but were designed to be interesting and distract attention away from the campaign material. Also, within both dynamic information boards were additional attention checks – pop up boxes that asked subjects to close the box within 10 seconds. In total, the dynamic information boards presented much more information than a person could possibly access, process and use in a single session. In such a setting, we expect than any differences caused by pay rates should become clearly observable.

Results of the Survey Experiment of Patriotism

Our first study was a survey experiment asking American citizens about how they felt about their state and nation. Subjects first viewed a randomly selected picture designed to elicit varying levels of national patriotism: a picture of Grand Central station (which to most observers unfamiliar with New York City looked simply like a large atrium with people walking through it), that identical image with a small digitally-inserted American flag in plain view, or a large image of the American flag. Subjects then were asked about how much they agreed with statements about being proud of the United States, the United States being the greatest nation in the world, and whether people born abroad could ever truly understand what it is to be an American. Subjects also were asked how long they have been a citizen of the United States. Next, subjects were queried about what state they lived in, and the previous questions were then repeated for their state. Subjects were then asked the "Moreno" question about whether they felt they were more a citizen of their country or of their state.[11] Finally, subjects were asked about their age, sex, partisanship and some political knowledge questions about their state's government. This provides several measures of who signed up at each pay rate, and how they participated within the study generally.

The first area where pay rates may have influenced participation in our studies was in self-selection of subjects choosing to accept and complete the HIT. We included only a few demographic questions in the first study, as we wanted to keep it as short as possible and we had no a priori reason to suppose demographic differences in the sample. What we found suggests little influence of pay rates upon sample composition.

<<<< Insert Table 1 here >>>>

We found no significant differences in the composition of our sample based upon age, sex or partisanship (although there is a consistent, though not significant, trend of fewer women as pay rates got higher). It is possible that the lack of differences found here was caused by these questions appearing at the end of the study, rather than at the beginning. It could be that the pay rate groups began

[11] The question was developed by social scientist Luis Moreno in 1986 in studying Scottish identity, but has been explored further in recent work, by among others McCrone and Bechhofer 2015. The question asks respondents to choose how salient their national identity is compared to other regional identities, in this case state identities.

demographically imbalanced, but only those subjects who felt adequately compensated successfully completed the study, and that it is only the reduced sample that showed no demographic differences. However, we have no evidence of this and find no reason to suspect this was the case. As a check, in our second study we placed the demographic questions at the beginning of the study.

<<<< Insert Table 2 here >>>>

A second area where subjects could have demonstrated pay-based differences was in their participation in the study. We specifically look at how many subjects in each pay group completed the study, how long they spent in the study (an indication of their attention), how many of the political knowledge questions they correctly answered, and how much time they spent answering those questions (measures of their effort in answering correctly). Along three of these measures (total time, time spent on political knowledge questions, and number of correct answers), subjects in the various pay groups exhibited no significant differences.

The only area we found significant differences based upon our rate of payment was in the completion rate of subjects.[12] Only about 70% of subjects who signed in from the lowest payment group (15 cents) completed the study, while approximately 85% of the highest paid subjects successfully completed the study. The two middle payment groups had completion rates around 80%. This data comes from the Qualtrics survey, which could only be accessed by Turkers who had

[12] With this number of dependent variables, it would be appropriate to correct our results for multiple hypothesis testing. However, given the pattern of a lack of results throughout our analysis, we are confident in our assertion that there are no effects and include the Holms corrections in the Online Appendix.

accepted the HIT. Since we only paid those who completed the HIT and returned the correct completion code, this means that some Turkers started the HIT and abandoned it, allowing others to then accept the HIT and complete it.

This suggests that subjects who feel underpaid may simply end their participation outright, rather than remaining in the study with inattentive performance. While this is not ideal, it is the better of the alternatives and seems to be a sort of self-selection of providing missing data from subjects who feel they are not being sufficiently compensated.

Table 1 and 2 demonstrate that pay rates in the survey experiment did not seem to effect who accepted the HIT or how they performed during the study. We can also examine how subjects answered the subjective questions in the study for any discrepancies there. While we have no theoretical reasons to believe that people accepting different compensation would differ across our measures, it is possible that people who feel differently about their compensation during the study might answer differently based upon their attention to or consideration of the questions. Subjects who feel undercompensated may simply speed through the questions, clicking arbitrarily without reference to the subject matter. If this were the case, we would expect this to occur primarily in the lower paid conditions, and less so in the higher paid conditions. We find no evidence of this however.

<<<< Insert Table 3 here >>>>

Table 3 presents the mean scores for the national and state level variable patriotism questions in our study, including the duration of residence question. Our four pay groups were statistically indistinguishable based upon their declared feelings of pride in, the greatness of, and the difficulty of being a full citizen in both the nation and their state. They also did not differ on the balance they felt between their national and state citizenship, which tilted towards greater feelings of national citizenship. In all, we find remarkably few differences across a range of measures in this study, and can only conclude that pay rates have very little influence on subject behavior in survey experiments.

Online Appendix Tables

Online Table 1. Subject Demographics of the Survey Experiment, by Pay Rate

Pay Rate	% Female	% Democrat	% Independent	Length of Residence in State	Length of Residence in USA	Mean Age Group
$0.15 (n=355)	42.6%	48.7%	15.6%	0.59 (0.02)	0.84 (0.02)	4.34 (0.12)
$0.30 (n=302)	41.0%	54.1%	13.9%	0.63 (0.02)	0.90 (0.02)	4.20 (0.14)
$0.50 (n=337)	40.5%	57.7%	14.4%	0.64 (0.02)	0.92 (0.01)	4.57 (0.13)
$0.75 (n=317)	35.0%	55.3%	13.7%	0.63 (0.02)	0.92 (0.01)	4.25 (0.12)
Total (n=1311)	39.8%	53.9%	14.4%	0.62 (0.01)	0.89 (0.01)	4.34 (0.06)
Pearson Chi^2	4.149	5.356	0.551			
F Statistic				1.139	6.808***	1.613

Online Table 2. Subject Participation in the Survey Experiment, by Pay Rate

Pay Rate	% Finished	Mean Seconds To Finish	Mean Seconds in Pol. Know.	Mean Correct Pol. Know.
$0.15 (n=355)	71.3%	195.45 (8.17)	56.78 (3.41)	2.85 (0.10)
$0.30 (n=302)	81.8%	173.38 (6.22)	49.09 (2.77)	3.04 (0.09)
$0.50 (n=337)	77.4%	181.67 (6.48)	52.78 (3.72)	2.97 (0.10)
$0.75 (n=317)	85.8%	191.59 (6.45)	54.64 (2.90)	3.02 (0.09)
Total (n=1311)	78.8%	185.68 (3.44)	53.46 (1.62)	2.96 (0.05)
Pearson Chi^2	23.345***			
F Statistic		2.057	0.990	0.831

Online Table 3. Subject Responses in the Survey Experiment, by Pay Rate

	State Patriotism			National Patriotism			Balance
Pay Rate	Pride	Great-ness	Citizen	Pride	Great-ness	Citizen	Moreno Question
$0.15 (n=316)	0.70 (0.02)	0.50 (0.02)	0.43 (0.02)	0.77 (0.01)	0.66 (0.02)	0.40 (0.02)	-0.09 (0.01)
$0.30 (n=266)	0.69 (0.02)	0.48 (0.02)	0.40 (0.02)	0.75 (0.01)	0.63 (0.02)	0.37 (0.02)	-0.09 (0.01)
$0.50 (n=293)	0.70 (0.02)	0.47 (0.02)	0.42 (0.02)	0.77 (0.01)	0.65 (0.02)	0.39 (0.02)	-0.10 (0.01)
$0.75 (n=301)	0.67 (0.02)	0.44 (0.02)	0.39 (0.02)	0.76 (0.01)	0.63 (0.02)	0.36 (0.02)	-0.10 (0.01)
Total (n=1176)	0.69 (0.01)	0.47 (0.01)	0.41 (0.01)	0.76 (0.01)	0.64 (0.01)	0.38 (0.02)	-0.10 (0.01)
F Stat	0.699	1.609	1.025	0.385	0.646	0.752	0.586