Research Article

Research and Politics
July-September 2018: 1–8
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053168018785483
journals.sagepub.com/home/rap
SAGE

Taking the time? Explaining effortful participation among low-cost online survey participants

Ian G. Anson

Abstract

Recent research has shown that Amazon MTurk workers exhibit substantially more effort and attention than respondents in student samples when participating in survey experiments. In this paper, I examine when and why low-cost online survey participants provide effortful responses to survey experiments in political science. I compare novice and veteran MTurk workers to participants in Qualtrics's qBus, a comparable online omnibus program. The results show that MTurk platform participation is associated with substantially greater effort across a variety of indicators of effort relative to demographically-matched peers. This effect endures even when compensating for the amount of survey experience accumulated by respondents, suggesting that MTurk workers may be especially motivated due to an understudied self-selection mechanism. Together, the findings suggest that novice and veteran MTurk workers alike are preferable to comparable convenience sample participants when performing complex tasks.

Keywords

MTurk, qBus, matching, satisficing, survey experiments

Introduction

Political scientists have recently debated whether MTurk workers' behavior differs from that of respondents recruited using other means. MTurk workers are suspected of exhibiting greater *compliance* and *attentiveness* than undergraduates and other online participants when performing tasks. Especially frequent MTurk participants—so-called "professional Turkers"—are thought to be extremely attentive in order to maximize their chances of receiving payment (Hauser and Schwarz, 2016). This behavior is believed to occur in part because many veteran MTurk workers have completed hundreds of Human Intelligence Tasks (HITs)[1] containing "attention checks": questions that measure respondents' levels of engagement with survey content (Hauser and Schwarz, 2016; Hillygus et al., 2014; Krupnikov and Levine, 2014; Mullinix et al., 2015).

Despite these concerns, the use of online respondents as experimental subjects in political science continues to grow in popularity. Some studies have shown that the unrepresentative nature of these samples can be easily improved through the use of conventional weighting strategies and screening procedures (e.g., Berinsky et al., 2012; Huff and Tingley, 2015; Levay et al., 2016; Thomas and Clifford, 2017). This literature also provides some evidence that experimental effect sizes in MTurk experiments are comparable to representative national surveys (but see Krupnikov and Levine, 2014).

However, MTurk is no longer the only option for researchers hoping to conduct low-cost survey experiments. In response to the growing demand for such data, new platforms for survey deployment have also recently arisen. One notable platform increasingly used by social scientists has recently been launched by Qualtrics, Inc. This platform, called "qBus" to indicate its function as a survey omnibus, enables low-cost access to national convenience samples. These samples feature demographic profiles that

University of Maryland, Baltimore County, MD, USA

Corresponding author:
Ian G. Anson, Department of Political Science, University of Maryland, Baltimore County, 1000 Hilltop Cir., 305 PUP, Baltimore,
MD 21250, USA.
Email: iganson@umbc.edu

approximate Census estimates of age, income, gender and race. At the time of this writing, the cost of a simple survey experiment on a qBus module is not substantially more than the deployment of a survey on MTurk. This is because many basic demographic questions that would otherwise extend the length of an MTurk survey are included for free as a part of the qBus omnibus.[2]

Existing research on low-cost survey participants—specifically MTurk workers—demonstrates high levels of attentiveness relative to student samples (e.g., Hauser and Schwarz, 2016). However, current research has been unable to distinguish whether effort differences arise because MTurk samples are unrepresentative or because the MTurk platform incentivizes high levels of attention. Further, no study has yet compared the effort levels observed across low-cost platforms. An inquiry in this vein can shed light on the ways in which the incentive structures of (and participation in) different online platforms can influence respondent behavior.

The present study has three objectives. First, I introduce the Qualtrics qBus platform and compare it to Amazon AWS's MTurk. Second, I examine the effects of platform type and frequency of survey participation on the effort exerted by low-cost online survey participants. Finally, on the basis of these findings, I assess the utility of Amazon MTurk and the Qualtrics qBus platform for contemporary practitioners.

Drawing on data from two samples of MTurk workers and two national online convenience samples of participants recruited by Qualtrics, Inc., I use calipered genetic matching techniques to compare respondent effort across platform and participation frequency. The results demonstrate that MTurk workers provide greatly increased attention check accuracy and content quantity relative to demographically-matched Qualtrics panelists. They also demonstrate that highly-active workers show virtually no differences in effort relative to participants with low participation rates. In a concluding section, I argue that a self-selection mechanism may be at play among online survey participants. This selection mechanism is likely the product of differing incentives and recruitment methods across the two platforms. While MTurk participants sign up as workers in order to earn cash payments, qBus respondents are recruited via social media accounts and receive cash equivalents rather than cash as payment.

Finally, the results also show that participants in qBus studies exhibit surprisingly low levels of effort regardless of prior survey experience. Relative to MTurk workers, participants on qBus platforms fail Instructional Manipulation Checks (IMCs) at a much higher rate. Further, these participants exerted less effort on an explicitly political task, meaning that qBus samples may be less appropriate than MTurk samples for political science studies measuring subtle experimental treatments.

MTurk and Qualtrics qBus: a preliminary comparison

Existing literature on MTurk has effectively described the processes of recruitment, payment and participation on that platform (e.g., Berinsky et al., 2012). However, no study has detailed the survey experience among Qualtrics participants. According to Qualtrics' European Society for Opinion and Marketing Research (ESOMAR) documentation, qBus participants are recruited using 'traditional, actively managed market research panels' (Qualtrics, 2014: 3) in addition to social media recruitment methods. Much like Survey Sampling International (SSI), another firm that provides access to survey participants, the third-party panels used by Qualtrics have been certified for quality by Mktg Inc.'s Grand Mean Certification Program. A list of Grand Mean certified survey firms is available at www.mktginc.com.

Qualtrics selects adult panelists residing in the US to participate in surveys. They are sampled on the basis of demographic quotas; this demographic information is collected through initial screening surveys. Panelists are contacted by Qualtrics via email or social media accounts, inviting them to participate in a survey for research purposes only. The Qualtrics documentation states that respondents' rewards for survey participation include "airline miles, gift cards, redeemable points, sweepstakes entrance and vouchers" (Qualtrics, 2014: 5). In addition, Qualtrics seeks to limit frequency of participation by ensuring that historical records are maintained for each panelist. This reduces the ability of qBus participants to act like "professional panelists" as the number of invitations they receive is necessarily limited.

qBus omnibus surveys normally include questions from three to five different researchers, most of whom are marketing researchers. These surveys include a battery of basic demographic questions, which can be modified to include political questions like partisanship and ideology. Each researcher's question block is presented to respondents in random order.

Importantly, Qualtrics makes a concerted effort to monitor and safeguard against respondent inattentiveness and lack of effort. Unlike the MTurk platform, qBus removes respondents who exhibit evidence of extreme "speeding" behavior (answering survey questions too quickly). However, given the high levels of attentiveness exhibited by MTurk workers, it is currently unclear whether these strategies help qBus samples to outperform MTurk samples in terms of average effort.

Explaining effort in the survey response

Individual survey respondents vary in the amount of effort they are willing to devote to survey participation. As respondents engage with a survey task, their cognitive resources deplete, yielding decreased effort (Krosnick, 1991; Narayan and Krosnick 1996). In the present study I consider two operationalizations of this so-called "satisficing" behavior. One is the failure of Instructional Manipulation Checks (IMCs). These tasks often tap respondents' attentiveness by asking them to give answers that would be uncommonly selected if not for seemingly inconsequential directions (Oppenheimer et al., 2009; Lelkes et al., 2012). The *quantity* of responses to open-ended questions also serves as a second relevant operationalization (Cerasoli et al., 2014).

Unpacking the "MTurk effect"

Despite recent advances, existing research on effort in low-cost survey samples has not clearly distinguished the effects of several determinants. These include the effects of platform experience, respondent self-selection and cross-sample demographic differences. Existing research has compared MTurk workers to college students, who do not demographically approximate the average MTurk sample (Hauser and Schwarz, 2016). This is especially problematic because effort is conditioned by individual-level attributes such as education (e.g., Krosnick, 1991).

Further, it is currently unclear if MTurk workers are more effortful because they have *learned* to exhibit such behavior through their experience on the platform. Some scholars have asserted an "MTurk effort thesis", which states that workers' prior experience on the MTurk platform causes them to exert high levels of effort (Berinsky et al., 2012; Berinsky et al., 2014; Brüggen and Dholakia, 2010; Goodman et al., 2013; Oppenheimer et al., 2009). Lastly, we do not know if effort is influenced by self-selection effects derived from platform incentives and recruitment strategies.

Sources of cross-platform variation in effort

One way to think about sources of variation in respondent effort is from the perspective of the Neyman–Rubin potential outcomes framework (e.g., Imai et al., 2008; Morgan and Winship, 2007; Rubin, 2005). From this perspective, differences in respondent effort levels could stem from the experiences of individuals who have been "treated" to the MTurk platform relative to members of the population who have instead experienced the qBus platform. The effect of platform experience is therefore the average treatment effect or ATE, specified as follows:

$$ATE = \frac{1}{N}\sum_{i=1}^{N}(Y_i\left(MTurk\right) - Y_i\left(qBus\right))$$

The measurement of the "true" ATE is theoretical, given that Y_i is observable in only one of the two conditions. As a result, estimation of the "MTurk effect" (ATE) relies on a comparison of different groups of treated and untreated units. However, as mentioned above, these estimates are confounded by demographic imbalance and self-selection. Differences in the effort levels of samples can be represented by $\widehat{ATE} = ATE - \Delta$ (Imai et al., 2008). Differences between MTurk and qBus respondents' effort levels are therefore attributable to one of three determinants: ATE, the true effect of the *experience* of MTurk participation on behavior; Δ_S, the effects of self-selection into platform participation; and Δ_T, effects that arise from demographic "treatment imbalance" across the groups.[3]

Effort and self-selection

Among these three components, I anticipate that self-selection is the most important driver of cross-platform differences in respondent effort. While many commentators have recently voiced concerns about the existence of savvy "professional Turkers" who discuss strategies for low-effort responses on forums, such shirking strategies are probably rare. On the qBus platform, workers are recruited via social media accounts, and are compensated in the form of gift cards and other cash-like forms of payment. MTurk workers are paid in dollar amounts to complete surveys, and are not recruited by social media invitation. These differences in recruitment method and payment method are likely to mean that MTurk workers possess greater intrinsic effort. This assertion stands in contrast to more conventional learning-based theories, which argue that online respondents become increasingly effortful as they learn to avoid attention checks.

Given the observational nature of the present study, blocking on demographic covariates can help to compensate for the effects of Δ_T, or treatment imbalance on the basis of observables. While we cannot fully disentangle the effects of survey experience from self-selection in this study, we can measure the effects of increased survey participation rates on effort levels. If we observe few differences in respondents' effort levels across survey participation rates, Δ_S, the effect of self-selection is likely to be substantial.

The above discussion leads to the following set of expectations:

> H1: Among online survey respondents, participation on the MTurk platform will be associated with a higher likelihood of IMC success than participation in the qBus online omnibus program.

> H2: Among online survey respondents, increased survey experience will not be associated with changes in the likelihood of IMC success.

> H3: Among online survey respondents, participation on the MTurk platform will be associated with increased response

Table 1. Results of matching analysis comparing IMC success rates, MTurk and qBus samples.

	qBus ATE	t	*p*-value
t-test	−0.197	−7.988	$p < 0.001$
Matched est.	−0.203	−8.05	$p < 0.001$

Note: Single nearest-neighbor genetic matching, caliper(1). Treated N = 503, matched N = 2446.

quantity compared to participation in the qBus online omnibus program.

Research design

In 2017 I performed a series of experiments (hereafter Study 1) which measured respondents' self-perceptions of political knowledge before and after exposure to various experimental treatments. Studies of identical design were fielded on a Qualtrics qBus omnibus (overall N = 1,047; present study N = 503) in May 2017 and through the MTurk platform (N = 1,559) in June 2017. Following existing work which compares MTurk samples to other samples, an HIT completion rate of 90 percent or higher was required (Hauser and Schwarz, 2016). The qBus omnibus is a national sample of respondents selected in a representative fashion on the basis of Census percentages for age, gender, ethnicity, household income and region. Qualtrics recruits respondents using actively-managed social media recruitment tools and other sources, according to their official documentation. qBus studies include pre-test demographic question batteries. The qbus module designed by the researcher included an IMC for 503 randomly-selected respondents, along with a question that asked respondents how many surveys they had completed in the past week. The IMC asked respondents to rank four objects from largest to smallest. While IMCs may be included on other omnibus users' question batteries, it does not appear that Qualtrics includes such tasks on qBus instruments itself. However, an initial survey question asks respondents to agree to provide responses that reflect their best effort. See the Supplementary Information (hereafter SI) for more details about the surveys, including demographic comparisons and question wording.

In 2016, another pair of experiments (hereafter Study 2) examined satisficing in an explicitly political context. The experiment was designed to measure priming effects on the accuracy of respondents' electoral forecasts. See Anson (n.d.) for a complete description of the theoretical approach. In August 2016, I reached a sample of 1,502 respondents through Amazon MTurk, and randomly exposed 512 of them to a treatment that contained an open-ended question asking why they felt a Presidential candidate would win the election. This survey also included an additional set of demographic and political questions. Then, in late October 2016, I reached a sample of 1,046 online respondents through a qBus omnibus survey administered by Qualtrics, 341 of whom received the writing task treatment.[4] Due to space constraints, please see the SI for a fuller description of this second study. In Study 1, effort was captured by a binary measurement that assessed the successful completion of a standard IMC task. In Study 2, effort was measured by the length of text provided in response to an open-ended political question.

Methods

In order to distinguish platform effects from demographic effects, I rely on matching techniques pioneered by Diamond and Sekhon (2013) to produce matched average treatment effect (ATE) estimates of the "treatment" of MTurk panel participation on the dependent variables in the study. Covariates including age, gender, education level, race, income and party identification were used in both studies to achieve matching. In Study 1 we also benefit from the inclusion of a measure of political knowledge. See the SI for sensitivity analyses that demonstrate that the present findings are likely robust to the influence of powerful unobservables (e.g., DiPrete and Gangl, 2004).

For each dependent variable of interest, I present a genetic matching analysis that relies on weights computed by the "GenMatch" software for the R programming environment (Sekhon, 2011). In the SI, additional models are presented as robustness checks, alongside evidence of match balance. The models presented below employ caliper boundaries of 1 standard deviation to sufficiently exclude pairs with low common support.

Study 1

Initial results from Study 1 are presented in Table 1. They demonstrate that matched MTurk respondents were much more likely to pass an IMC than their demographically-similar qBus counterparts. It appears that matched respondents in the MTurk sample were roughly 20.3% more likely than the qBus respondents to successfully complete the IMC, even after matching on demographics and short-term survey participation rates. This difference is striking, as it reflects an average increase in accuracy from around 55% to around 75%. Compensating for demography and the frequency of survey participation, qBus respondents were relatively poor performers on the IMC, which was not a particularly difficult task. Respondents were asked to rank objects from smallest to largest, including a pineapple, a tree, a mouse and a pea. The fact that almost half of the qBus participants failed this task is an alarming indication of the overall level of effort in this sample.

We next turn our attention to the reasons for variation within and across the samples. Table 2 presents the results of logistic regression models predicting effort. These

Table 2. Logistic regression models predicting likelihood of IMC success, Study 1.

	Model 1	Model 2	Model 3
	Combined	MTurk sample	qBus sample
qBus respondent	−0.694***		
	(0.131)		
Survey participation rate	−0.00005	−0.0001	0.001
	(0.001)	(0.001)	(0.003)
Male	−0.203	−0.180	−0.194
	(0.104)	(0.125)	(0.203)
Nonwhite	−0.153	−0.238	−0.033
	(0.116)	(0.142)	(0.213)
Democrat	−0.312**	−0.239	0.525
	(0.118)	(0.140)	(0.222)
Republican	−0.296*	−0.222	−0.459
	(0.141)	(0.173)	(0.259)
Income	−0.208**	−0.135	−0.333*
	(0.073)	(0.090)	(0.135)
Education	0.025	0.035	0.018
	(0.039)	(0.050)	(0.065)
Age	−0.002	−0.005	0.003
	(0.004)	(0.006)	(0.006)
Political knowledge	0.235***	0.238***	0.230**
	(0.044)	(0.053)	(0.079)
qBus resp. x survey rate	0.001		
	(0.003)		
Constant	1.030***	0.928**	0.515
	(0.240)	(0.321)	(0.392)
N	2013	1510	503
Log likelihood	−1154.996	−820.115	−332.342
AIC	2333.992	1660.23	684.685

Note: Standard errors in parentheses. ***p < 0.001; **p < 0.01; *p < 0.05.

models allow us to estimate the interaction between survey mode and the number of surveys that respondents reported completing in the past week.

Table 2 presents models predicting IMC success for the combined samples (Model 1), the MTurk sample alone (Model 2) and the qBus sample (Model 3). The results in the ‘Model 1’ column demonstrate that the frequency of survey completion has almost no effect on IMC accuracy when considering the full sample. In addition, this model shows that an interaction between MTurk participation and the rate of survey participation also has virtually no distinguishable effect—an indication that more experienced MTurk and qBus workers are no more or less likely than relatively inexperienced respondents to fail IMCs. These results hold for Models 2 and 3, respectively, which examine the two surveys separately.

The only theoretically-relevant significant effect across the models is the baseline difference between MTurk and qBus IMC success rates. According to Model 1, qBus respondents are around 2.06 times more likely than MTurk respondents to fail the IMC ($p < 0.001$; an effect size that is somewhat larger than that seen in the matched results above in Table 1). However, this result provides a robustness check that works to confirm the wide gap in IMC attentiveness across survey platforms, net of relevant demographics.

Study 1 also allows us to assess H2. These results show the relatively small impact of “professional” online survey participation status on IMC success rate across both qBus and MTurk samples. Figure 1 shows this relationship in more fine-grained detail, through a Loess-smoothed plot of survey participation rates on IMC success rate. The plot shows the predicted success rate with a 95% confidence interval represented by the shaded area.

The results once again confirm that very frequent survey participants are no better or worse than average or even very infrequent survey-takers, despite a slight (though statistically non-significant) positive slope for qBus participants across the range of the weekly survey completion variable ($\beta_{MTurk} = -0.00001, p = 0.98$; $\beta_{qBus} = 0.001, p = 0.79$). The major differences to emerge in the study thus far are instead found when comparing respondents across the platforms. Low levels of effort among qBus respondents are made clear by Figure 1: the predicted likelihood of IMC success for MTurk workers at any point in the graph stands above 70%, whereas even the most savvy qBus participants’ success rates are closer to 50% on average.

Thus far, we have observed evidence in support of H1, which argues that MTurk workers will exhibit greater attentiveness than qBus panelists. We also observe evidence in support of H2, which states that survey participation rates will not affect effort levels. It is still unclear whether cross-platform differences are unique to IMC tasks, however, as MTurk workers may quickly learn to identify these items through early exposure to the platform.

Study 2

In Study 2, I assess H3 by examining whether MTurk respondents provided more detailed written responses than qBus respondents to a substantive, open-ended political question. This question is unlikely to be interpreted as an IMC. In this open-ended response, participants were asked to explain their expectations regarding the eventual outcome of the 2016 Presidential election. Table 3 presents the results of a calipered genetic matching comparison of logged response length to this political question.

The results of this additional analysis demonstrate that, again, qBus respondents were substantially less effortful in their participation in the open-ended political item than were MTurk workers. It appears that matched respondents

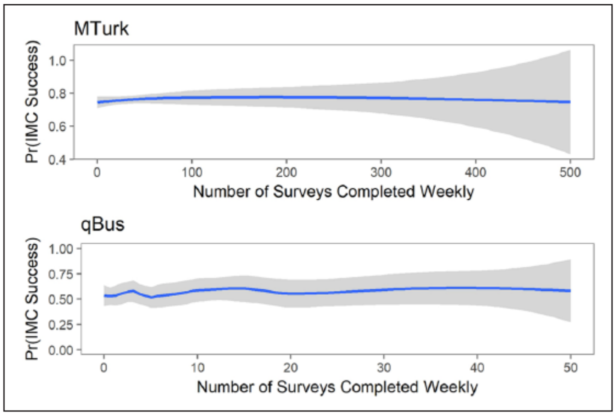

Figure 1. Effect of survey participation rates on IMC success, Study 1.

Table 3. Results of matching analysis comparing number of characters written (logged), MTurk and qBus samples.

	qBus ATE	*t*	*p*-value
t-test	−0.541	−9.102	$p < 0.001$
Matched est.	−0.561	−8.52	$p < 0.001$

Note: Single nearest-neighbor genetic matching, caliper(1). Treated $N = 340$, matched $N = 600$.

in the MTurk sample wrote around 121 characters on average, compared to an average response length of roughly 68 characters for the qBus sample. This difference of around 53 characters represents the length of an additional short sentence. In addition to the successful completion of IMCs, which even novice Turkers may come to quickly recognize, the effort devoted to an open-ended substantive task shows evidence of increased effort among MTurk workers relative to qBus respondents.

Conclusions

At first glance, the present findings work to confirm a suspected pattern: MTurk respondents proffer a greater amount of effort than comparable online survey panelists. This pattern holds for open-ended responses and IMC tasks. Based on these findings alone, the "MTurk effect" thesis finds support: after compensating for demographic differences, MTurk respondents still provide greater effort on a variety of tasks than workers on the qBus platform.

But in contrast to the conventional wisdom regarding the "MTurk effect", the results also show that the frequency of participation on online survey platforms has little effect on effort levels. Both qBus and MTurk respondents with high amounts of survey participation are no less effortful than participants with almost no recent survey experience on each platform. Scholarship is poised to further investigate why this is the case. The existing literature on IMCs suggests that both qBus and MTurk respondents should exhibit increased attentiveness and successful IMC completion rates as they gain experience in the survey setting. However, as this pattern is not supported by the data, the ATE estimates observed in Studies 1 and 2 are more likely caused by strong self-selection effects. As recruitment and payment methods differ across the two platforms, respondents are differentially motivated to provide their best efforts when engaging with the survey task.

This study possesses several important limitations. Perhaps most importantly, general inferences about the nature of "MTurk workers" or "qBus panelists" writ large are dubious when relying on data from just four samples—despite recent evidence that MTurk samples are more stable in composition than previously assumed (Clifford et al., 2015; Shapiro et al., 2013). Additional shortcomings include the relatively small sample sizes and the limited number of demographic variables observed in both surveys. And while the number of surveys per week is a useful indicator of survey takers' intensity of platform usage, this variable does not fully capture respondents' historical usage of a platform.[5]

However, these findings do provide us with several broad takeaways for researchers hoping to perform low-cost survey experiments. First, we now know that increased survey participation among MTurk workers does little to influence attention levels. MTurk workers are notably effortful regardless of the inclusion of IMCs in a given study, and regardless of whether or not they have recently completed hundreds of surveys on the platform. We also observe very low baseline effort levels among qBus panelists, a finding that detracts from this service's appeal for contemporary experimental research. Finally, the results show that increasing the representativeness of MTurk samples, à la Levay et al. (2016), represents a promising way forward for political scientists seeking to perform low-cost survey experimental research.[6]

Acknowledgements

The author wishes to thank Carolyn Forestiere, William Blake, Simon Stacey, Tyson King-Meadows, Thomas Schaller, Marc Hetherington, Travis Braidwood, D. Sunshine Hillygus and the anonymous reviewers for their helpful comments on earlier versions of this paper.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplementary material

The supplementary files are available at http://journals.sagepub.com/doi/suppl/10.1177/2053168018785483

The replication files are available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YZJXUH

Notes

1. Amazon MTurk uses the term HIT to refer to any task performed by MTurk workers, including surveys.
2. The present study presents the results of two MTurk and two qBus survey deployments. While qBus generally advertises a rate of around $250 per question, the cost of the qBus deployments exceeded the analogous MTurk deployments in each case only by around $400 and $500, respectively, given that our MTurk respondents were paid for their time at a rate consistent with the statewide minimum wage in — (redacted state). Another advantage of the qBus program is the fact that surveys created using Qualtrics's online platform can be accessed directly by Qualtrics employees through the online interface. This means that users' survey designs are copied directly into the qBus module, guaranteeing their accuracy.
3. Decomposing Δ, $\Delta = \Delta_S + \Delta_T, = \Delta_{S_X} + \Delta_{S_U} + \Delta_{T_X} + \Delta_{T_U}$ where Δ_{S_X} corresponds to sample selection bias based on observables; Δ_{S_U} corresponds to sample selection bias based on unobservables; and Δ_{T_X} and Δ_{T_U} signify treatment imbalance based on observables and unobservables, respectively.
4. As these groups were properly randomized, we should not expect this decision to affect the results.
5. I thank an anonymous reviewer for this point.
6. In the Appendix, I use post-stratification weighting of the MTurk surveys to approximate the demographic profile of the qBus surveys. Results show that after weighting, cross-sample comparisons of effort are consistent with the results seen above. This supplemental analysis shows that the approach used by Levay et al. (2016) is preferable to the use of qBus panels for experimental research.

ORCID iD

Ian G. Anson https://orcid.org/0000-0002-1545-4270

References

Anson IG (n.d.) Partisan cheerleading in electoral forecasts. Working paper. Available at: www.iananson.com/papers

Berinsky AJ, Huber GA and Lenz GS (2012) Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* 20(3): 351–368.

Berinsky AJ, Margolis MF and Sances MW (2014) Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58(3): 739–753.

Brüggen E and Dholakia UM (2010) Determinants of participation and response effort in Web panel surveys. *Journal of Interactive Marketing* 24(3): 239–250.

Cerasoli CP, Nicklin JM and Ford MT (2014) Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin* 140(4): 980.

Clifford S, Jewell RM and Waggoner PD (2015) Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics* 2(4). DOI: 10.1177/2053168015622072.

Diamond A and Sekhon JS (2013) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95(3): 932–45.

DiPrete TA and Gangl M (2004) Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology* 34(1): 271–310.

Goodman JK, Cryder CE and Cheema A (2013) Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26(3): 213–224.

Hauser DJ and Norbert Schwarz (2016) Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* 48(1): 400–407.

Hillygus DS, Jackson N and Young M (2014) Professional respondents in non-probability online panels. In: Callegaro M, Baker R, Bethlehem J, Goritz AS, Krosnick J and Lavrakas PJ (eds) *Online Panel Research: A Data Quality Perspective*. 1st edition. Chichester: Wiley, pp. 219–237.

Huff C and Tingley D (2015) "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics* 2(3). DOI: 10.1177/2053168015604648.

Imai K, King G and Stuart EA (2008) Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2): 481–502.

Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5(3): 213–236.

Krupnikov Y and Levine AS (2014) Cross-sample comparisons and external validity. *Journal of Experimental Political Science* 1(1): 59–80.

Lelkes Y, Krosnick JA, Marx DM, Judd CM and Park B (2012) Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology* 48(6): 1291–1299.

Levay KE, Freese J and Druckman JN (2016) The demographic and political composition of Mechanical Turk samples. *SAGE Open* 6(1). DOI: 10.1177/2158244016636433.

Morgan SL and Winship C (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Mullinix KJ, Leeper TJ, Druckman JN and Freese J (2015) The generalizability of survey experiments. *Journal of Experimental Political Science* 2(2): 109–138.

Narayan S and Krosnick JA (1996) Education moderates some response effects in attitude measurement. *Public Opinion Quarterly* 60(1): 58–88.

Oppenheimer DM, Meyvis T and Davidenko N (2009) Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45(4): 867–872.

Qualtrics (2014) *ESOMAR 28: 28 Questions to Help Research Buyers of Online Samples*. Technical Report.

Rubin DB (2005) Causal Inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469): 322–331.

Sekhon JS (2011) Multivariate and propensity score matching software with automated balance optimization: The Matching Package for R. *Journal of Statistical Software* 42(7). DOI: 10.18637/jss.v042.i07.

Shapiro DN, Chandler J and Mueller PA (2013) Using Mechanical Turk to study clinical populations. *Clinical Psychological Science* 1(2): 213–220.

Thomas KA and Clifford S (2017) Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* 77: 184–197.