Behavior Research Methods (2021) 53:2591–2595
https://doi.org/10.3758/s13428-021-01588-4

Turking in the time of COVID

Antonio A. Arechar[1,2,3] · David G. Rand[2,4,5]

Accepted: 22 March 2021 / Published online: 7 May 2021

Abstract

On March 16, 2020, the US Government introduced strict social distancing protocols for the United States in an effort to stem the spread of the COVID-19 pandemic. This had an immediate major effect on the job market, with millions of Americans forced to find alternative ways to make a living from home. As online labor markets like Amazon Mechanical Turk (MTurk) play a major role in social science research, concerns have been raised that the pandemic may be reducing the diversity of subjects participating in experiments. Here, we investigate this possibility empirically. Specifically, we look at 15,539 responses gathered in 23 studies run on MTurk between February and July 2020, examining the distribution of gender, age, ethnicity, political preference, and analytic cognitive style. We find notable changes on some of the measures following the imposition of nationwide social distancing: participants are more likely to be less reflective (as measured by the Cognitive Reflection Test), and somewhat less likely to be white, Democrats (traditionally over-represented on MTurk), and experienced with MTurk. Most of these differences are explained by an influx of new participants into the MTurk subject pool who are more diverse and representative – but also less attentive – than previous MTurkers.

Keywords online research · COVID-19 · demographics · diversity · Amazon Mechanical Turk

Introduction

In March of 2020, the COVID-19 pandemic began to spread across the United States. On March 16, the US government announced new measures to curtail the virus, such as limiting travel, restricting public gatherings, and closing schools. This nation-wide social distancing policy – hereafter "quarantine" for simplicity – had an immediate impact on the job market, with millions of Americans suddenly unemployed.

Concerns have been raised that this change in the labor market would also affect the subject pool available on crowdsourcing marketplaces, such as Amazon Mechanical Turk (MTurk), which are cornerstones for subject recruitment across the experimental social sciences (Horton et al., 2011). For instance, subject pools may become less representative if people experiencing economic hardship due to the pandemic no longer have access to sufficiently reliable Internet connections to participate in online studies (Lourenco & Tasimi, 2020); remain unaltered (Moss et al., 2020); or in fact become more diverse as the need for other forms of employment might drive new workers to such marketplaces.

Indeed, the quarantine may be changing the pool of workers available to take part in academic studies in two ways: either by influencing which existing workers participate, or by causing an influx of new ones. Depending on their nature and prevalence, samples drawn by researchers might appear more or less representative than or as representative as before the pandemic started. Hence, in order to assess any potential change in sample compositions, it appears imperative to first disentangle both effects. However, studying the influx of new participants is hard because researcher groups traditionally only recruit restricted samples with previous experience and acquired reputation on platforms (Peer et al., 2014).

✉ Antonio A. Arechar
antonio@arechar.com

David G. Rand
drand@mit.edu

[1] Center for Research and Teaching in Economics, CIDE, Aguascalientes, Mexico

[2] Sloan School of Management, (E-62) Room 539, Massachusetts Institute of Technology, 30 Memorial Dr, Cambridge, MA 02138, USA

[3] Center for Decision Research and Experimental Economics, University of Nottingham, Nottingham, UK

[4] Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA

[5] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Here we consider the impact of the current pandemic on subject pool composition by looking at 15,539 responses given to studies conducted by our group on MTurk between February and July 2020. Moreover, by exploiting a recent change in our recruitment approach, we are better suited to attract new participants, to track them over time (Robinson et al., 2019) and, relative to usual recruitment restrictions, to anticipate eventual surges – as many of the new accounts considered would meet their criteria within a few weeks (Rand et al., 2014).

We look at changes in participant demographics and the overall quality of the data produced, which can have important implications for social science researchers assessing representativeness and generalizability of results, as well as for comparing results from studies run pre-quarantine versus during (or after) quarantine. Specifically, we examine the distribution of gender, age, ethnicity, political preferences, and analytic cognitive style (as measured by the Cognitive Reflection Test, CRT (Frederick, 2005)) before and after the social distancing measures were introduced on March 16. In addition, we distinguish specific trends by type of participants: either first recruited using traditional qualifications ("baseline"), or first recruited with no restrictions pre- or post-quarantine introduction ("pre" and "post", respectively).

We find important differences on all measures during quarantine when comparing *between* types. In particular, baseline participants are somewhat less diverse than unrestricted ones, as they tend to be more reflective, female, white, older, and in favor of the Democratic party; pre-quarantine participants are also more reflective and whiter than post-quarantine ones. When comparing *within* participants before and during the pandemic, we note no evidence of structural changes for the unrestricted sample, whereas for the returning baseline sample, we find it to be younger, more ethnically diverse, and less reflective.

At the same time, we note a potential trade-off between diversity and overall data quality. Responses by unrestricted participants are noisier than the ones provided by the baseline, whereas responses from post-quarantine participants are somewhat noisier than their pre-quarantine counterparts.

We conclude by suggesting that researchers take these changes in the MTurk subject pool – as well as the potential for changes in other variables that we did not measure – into account. This pandemic has changed the way we interact with each other, as it has for the people we often employ as primary input in research.

Methods

We first identify all workers taking part in our MTurk studies in 2020. Based on the number of tasks taken, their date, and the criteria used for initial recruitment, we then classify participants into three categories: (i) "baseline" workers who had at least 100 studies taken on the platform and at least 95% of them approved at the time of initial recruitment – which occurred before February 24, the date when we changed our recruitment policy and allowed everyone located in the US to participate; (ii) workers recruited with no restrictions "pre"-quarantine (i.e., between February 24 and March 15); and (iii) workers first recruited "post"-quarantine. We finally selected the 23 studies conducted since February 24 with relevant demographic data, unrestricted recruitment, and at least 100 participants in them to identify our target sample of 15,925 observations (see Table S1 for study details).

Each of the individual difference variables we analyze appears in at least 12 of the 23 studies surveyed (see Fig. S1 for distributions over time). Gender and ethnicity had different scales across studies; for comparability, we contrast male vs. female and white vs. non-white, respectively. Political preference appeared in some studies on a six-point Likert scale and in others on a seven-point one, so we set both as percentages with a minimum of 0 ("Strongly Democrat") and a maximum of 1 ("Strongly Republican").

Since we match records from two different datasets (one from Qualtrics to identify participants' choices, and one from MTurk to identify their category), we lose 386 (2%) observations where participants completed part of the study but did not submit it back to MTurk – leaving a final sample of 15,539 observations. Most of those entries have missing values in the individual difference variables but, among those with data, we find that non-complete respondents differ from complete ones in that they are represented by a less white and less reflective population (ps < 0.003).

Most, but not all, participants appear only once in the collated dataset (69% once, 22% twice, 6% three times, and the remaining 3% four times or more). Hence, all significance values reported are based on linear regressions with standard errors clustered on subject. Unless otherwise stated, dependent variables are the individual difference variables, whereas the independent variables are dummies for baseline and (unrestricted) pre-quarantine samples. In addition, we include a linear time trend to account for potential temporal patterns (see Fig. S1), but results remain qualitatively similar excluding it.

Results

We find a surge of brand-new participants since the COVID-19-related quarantine was announced. The share of new accounts per study was previously 17%, and it increased to 34% during quarantine. Relatedly, over 60% of participants in the last few weeks of the data gathered were recent and recruited unrestrictedly (Fig. 1a). It is thus apparent that the sample composition of our studies drastically changed during quarantine, mainly driven by new workers who joined after its announcement.

Are post-quarantine workers different? A comparison between the three types of workers outlined above reveals that, during quarantine, post-quarantine workers are more likely to be of non-white ethnicity and to perform more poorly on the CRT relative to baseline and pre-quarantine samples, and more likely to be younger, male, and favorable towards the Republican party relative to baseline; whereas baseline workers are more likely to be reflective, favorable towards the Democratic party, female, and older than the unrestricted samples recruited pre-quarantine (ps < .010; Fig. 1b and Table S2). Unrestricted pre- and post-quarantine samples are particularly closer to each other with regards to age and gender.

Moreover, when comparing workers before and during quarantine, we find no meaningful differences for the unrestricted ones. Yet, returning baseline workers are younger, less reflective, and less likely to be white (Table S3). Thus, we find converging evidence that samples during the pandemic tend to be somewhat more diverse but also less reflective, and that most of the differences observed are driven more by the influx of post-quarantine accounts than by changes in the returning ones.

Given the tendency for non-whites to disfavor the Republican party, the fact that so many new accounts identify as non-white Republicans raises the question of whether the apparent increase in diversity is simply the result of random responding by new workers (Chandler et al., 2020). Thus, we also explore how the data gathered from each type of participant fares in five quality checks during quarantine, and test if the differences seen in Fig. 1b. hold when excluding any potential outliers from the sample (for details see Table S4).

First, one might wonder if post-quarantine accounts are more likely to provide bot-like responses. We find pre- and post-quarantine accounts to be more likely to connect from a suspicious Internet service provider than baseline, and no difference between pre- and post-quarantine accounts (3% base; 14% pre; 10% post; $p < 0.001$ and $p = 0.068$; as classified by Ref. (Prims et al., 2018), which is updated periodically). Thus, we find some evidence that the quality produced by unrestricted accounts might be lower than baseline. When excluding suspicious accounts, all demographic differences reported in Fig. 1b remain significant though.

Second, we investigate whether new accounts rush through the tasks assigned by looking at the time taken to complete them (completion times are log-transformed due to right skew; study dummies are included). Post-quarantine participants actually take longer to complete studies relative to baseline but not to pre-quarantine ($p < 0.001$ and $p = 0.502$), suggesting that they are not rushing through questionnaires providing nonsense random answers. This is further confirmed when looking at deciles 1 and 10 of time taken per study, where the share of post-quarantine workers is lower and higher, respectively, than baseline workers (ps < 0.001) but not pre-quarantine ones (ps > 0.510). Thus, baseline differs from the other two samples in that they are faster completing studies overall. Removing outliers (deciles 1 and 10) from the analysis does not change the demographic differences depicted in Fig. 1b, apart from post-quarantine samples being relatively less white than the pre-quarantine ones.

Third, we look at the fraction of accounts completing studies from mobile devices, as it is possible that attention, and data quality overall, is influenced by the nature and functions of such devices. We find baseline working on mobile devices relatively more often than unrestricted samples (4% base; 2% pre and post), but the shares are so small that the general results in Fig. 1b barely change when excluding them.

Fourth, we evaluate subjects' consistency by comparing their responses in a follow-up study surveying age, gender, ethnicity, and political preference, but not CRT ($n = 744$; see Appendix C for details). In all, responses were fairly consistent (94%). Yet, we find baseline participants more likely to be consistent than post-quarantine workers (97% base; 91% pre; 88% post; $p < 0.001$), and no differences between pre- and post- or pre- and baseline accounts (ps > 0.067). Restricting to only consistent records, we find that the differences in Fig. 1b for political party and age remain significant for baseline, but not the ones reported for gender and ethnicity – although note that relative sample sizes for this particular check are rather small.

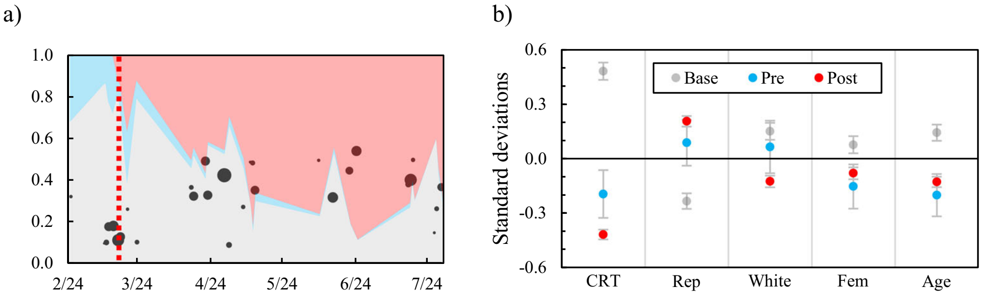

Fig. 1 **a** Share of workers per category over time. *Bubbles* represent the fraction and relative size of new accounts appearing in a given day. The *red dotted line* represents the day when quarantine was introduced. The areas in *gray*, *blue,* and *red* represent the fraction of baseline, pre-, and post-quarantine workers per day, excluding 4 days where sample size was less than 43 (i.e., 1, 1, 4, and 10). **b** Standardized mean differences between baseline, pre-, and post-quarantine workers. CRT: number of correct answers in the CRT; Rep: preference for the Republican party; White: white ethnicity; Fem: female; Age: worker age. 95% confidence intervals plotted

Finally, we evaluate attentiveness as a fifth quality check by looking at the CRT responses that were neither intuitive nor reflective, and at participants' performance in attention checks. We differentiate between CRT responses that were wrong because they gave the intuitive response versus other unexpected wrong answers, because people providing incorrect, non-intuitive answers may have failed to read the question or responded randomly – rather than merely failing to reflect upon their intuitive response. We find that post-quarantine workers give unexpected answers more often than the rest (*p*s < 0.001; Fig. 2a), but that the number of unexpected answers by new workers is rarely greater than 1 – that is, there is essentially no evidence of post-quarantine workers answering all items randomly (Fig. 2b).

With regards to performance on two attention checks (taken from Ref. (Berinsky et al., 2014)) administered in four of the studies run during quarantine (n = 2,407), unrestricted workers perform substantially worse on the attention checks overall but very few of them get both questions wrong (Fig. 2c). This again suggests inattention, but not completely random responding. Moreover, when comparing performance on a trivial attention check ("dog is to puppy as cat is to ___") collected as part of the follow-up survey reported earlier, we find post-quarantine accounts rarely fail to provide "kitten" as an answer, but that they nevertheless do so less often than baseline and at a similar rate than pre-quarantine accounts (100% base; 92% pre; 87% post; $p < 0.001$ and $p = 0.468$). Of note, removing the outliers found in this section does not change any of the differences depicted in Fig. 1b. Moreover, when excluding all but the outliers reported in the fourth check, because of the small sample, we only find two significance changes: relative to post-quarantine samples, pre-quarantine accounts are no longer whiter and become more favorable toward the Democratic party.

Discussion

Here we find that the sample composition of MTurk studies differs before versus during the COVID-19-induced national quarantine. New participants who entered during the quarantine are less reflective, and somewhat more likely to be non-white and Republican, even relative to other participants first recruited just a few weeks before. Indeed, these new participants are actually closer to the national average than baseline and pre-quarantine workers (most notably, scholars studying COVID-19 and related political issues using MTurk should note that, during quarantine, the traditional liberal bias of the MTurk subject pool can be practically eliminated, especially if conventional recruitment restrictions are removed). Thus, contrary to concerns that have been raised (Lourenco & Tasimi, 2020), quarantine has made MTurk more representative in numerous respects.

At the same time, there is some potential trade-off between representativeness and overall data quality, as post-quarantine workers are also more likely to be inconsistent, fail attention checks, and give responses on the CRT that are neither intuitive nor correct. To put the attentiveness of these new workers in context, it is useful to compare them to workers from other crowdsourcing platforms. To do so, we analyze large datasets collected recently using Prolific (Pennycook et al., 2021) and Lucid (Pennycook et al., 2020). Considering the one attention check question that was common to the studies run on all three platforms, we find a failure rate of 3% among baseline workers, 23% among pre- and post-quarantine workers, 21% among Prolific workers, and 75% among Lucid workers. The fraction of participants giving at least one unexpected CRT response is 15% among baseline workers, 29% among pre-quarantine workers, 38% among post-quarantine workers, 19% among Prolific workers, and 34% among Lucid workers. Furthermore, the fraction of MTurk workers with at least one unexpected response in a 2019 study with no restrictions other than being located in the US was of 25% (Robinson et al., 2019). Thus, new MTurk workers do not seem appreciably worse than participants from Lucid but they do seem to perform worse on the CRT than Prolific users and their MTurk counterparts a year ago.

One might wonder if our findings on attentiveness were influenced by overall lack of experience on the platform (Chandler et al., 2014), as new MTurk workers were not only less used to the demands of completing studies than more experienced ones but could also be less able to manage the competing demands on their time that quarantine brought with

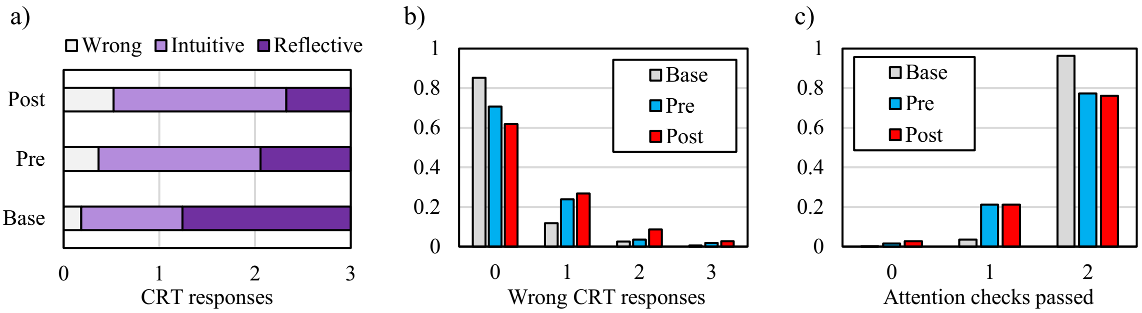

Fig. 2 **a** Distribution of responses in the CRT. **b** Distribution of responses in the CRT that are neither reflective nor intuitive. **c** Distribution of correct answers to two attention checks administered in four studies (Berinsky et al., 2014)

it. Although data on participants' total experience on the platform are not available, we can use our account's study count as a rough proxy and assess evolving performance in attention checks. Indeed, we note mild improvements with experience, but also that baseline workers perform better than unrestricted ones at any comparable level (Fig. S2). Furthermore, we find that including our proxy for experience as an additional control in two supplementary analyses (during pandemic and throughout 2020; Table S5), leaves our general findings consistently and qualitatively the same. Hence, although we cannot rule out a distinctive effect of (lack of) experience on new participants' performance, we find little support for it with our proxy.

Moreover, given that our recruitment has been unrestricted since February 24, it is possible that some workers first participating since then met the quality and quantity restrictions set in baseline, so the differences here reported between baseline and unrestricted (pre- and post-quarantine) samples might be underestimated.

Our results are seemingly in contrast with the one prior attempt we are aware of to investigate the COVID-19 pandemic's impact on the MTurk subject pool (Moss et al., 2020). This prior work used CloudResearch's publicly available Metrics tool to examine the demographics of nearly all MTurk subjects available on the platform, and found no differences in demographics due to the pandemic. Their work, however, did not specifically examine users participating in social science studies, which represent a notable minority of their participants (Litman et al., 2020). Moreover, the modal social science researcher using CloudResearch is likely to be applying the conventional quality restrictions we used in our baseline cohort. Thus, as the changes we observe are mainly driven by new workers, whereas our baseline workers display relatively more stable patterns over time, our findings are not necessarily in contrast with theirs.

In sum, our results suggest that there has actually been a meaningful change in social science participants, mainly driven by a steady influx of new participants. We hope these observations will be of use to researchers using MTurk, helping to allay concerns about demographic shifts while potentially raising attentiveness concerns. It seems likely that at least some of these changes will persist as the quarantine eases and many people are able to return to work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01588-4.

Acknowledgements We are grateful to Aaron Moss, Leib Litman, and the CloudResearch team for insightful conversations when comparing their findings with ours. We are also grateful to the anonymous reviewers, Jillian J. Jordan, Michael Stagnaro, Gordon Pennycook, Gordon Kraft-Todd, Adam Bear, and Arber Tasimi for helpful comments.

References

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (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.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, *129*(1), 49.

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

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, *14*(3), 399–425.

Litman, L., Robinson, J., Rosen, Z., Rosenzweig, C., Waxman, J., & Bates, L. M. (2020). The persistence of pay inequality: The gender pay gap in an anonymous online labor market. *PLoS One*, *15*(2), e0229383.

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends in Cognitive Sciences* https://doi.org/10.1016/j.tics.2020.05.003

Moss, A. J., Rosenzweig, C., Robinson, J., & Litman, L. (2020). Demographic stability on Mechanical Turk despite COVID-19. *Trends in Cognitive Sciences* https://doi.org/10.1016/j.tics.2020.05.014

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031.

Pennycook, G., McPhetres, J., Bago, B., & Rand, D. G. (2021). Beliefs about COVID-19 in Canada, the U.K., and the U.S.A. Retrieved from osf.io/3a497

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. *Psychological Science*. 31(7), 770-780.

Prims, J. P., Sisso, I., & Bai, H. (2018). Suspicious IP Online Flagging Tool. Retrieved from https://itaysisso.shinyapps.io/Bots

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., et al. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*(1), 1–12.

Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLoS One*, *14*(12), e0226394. https://doi.org/10.1371/journal.pone.0226394

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.