Journal of Experimental Social Psychology 70 (2017) 153–163

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Beyond the Turk: Alternative platforms for crowdsourcing behavioral research

Eyal Peer [a,*], Laura Brandimarte [b], Sonam Samat [c], Alessandro Acquisti [c]

[a] Graduate School of Business Administration, Bar-Ilan University, Ramat-Gan 52900, Israel
[b] Eller College of Management, University of Arizona, Tucson, AZ, United States
[c] Heinz College, Carnegie Mellon University, Pittsburgh, PA, United States

ARTICLE INFO

Article history:
Received 30 May 2016
Revised 19 January 2017
Accepted 21 January 2017
Available online 1 February 2017

Keywords:
Online research
Crowdsourcing
Data quality
Amazon Mechanical Turk
Prolific Academic
CrowdFlower

ABSTRACT

The success of Amazon Mechanical Turk (MTurk) as an online research platform has come at a price: MTurk has suffered from slowing rates of population replenishment, and growing participant non-naivety. Recently, a number of alternative platforms have emerged, offering capabilities similar to MTurk but providing access to new and more naïve populations. After surveying several options, we empirically examined two such platforms, CrowdFlower (CF) and Prolific Academic (ProA). In two studies, we found that participants on both platforms were more naïve and less dishonest compared to MTurk participants. Across the three platforms, CF provided the best response rate, but CF participants failed more attention-check questions and did not reproduce known effects replicated on ProA and MTurk. Moreover, ProA participants produced data quality that was higher than CF's and comparable to MTurk's. ProA and CF participants were also much more diverse than participants from MTurk.

In recent years, a growing number of researchers have used Amazon Mechanical Turk (MTurk), a crowdsourcing platform, to recruit online human subjects for research (Paolacci & Chandler, 2014). A large body of research has demonstrated that MTurk can be a reliable and cost-effective source of high-quality and representative data, for multiple research purposes, in and outside the behavioral sciences (e.g., Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Fort, Adda, & Cohen, 2011; Goodman, Cryder, & Cheema, 2013; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010; Rand, 2012; Simcox & Fiez, 2014; Sprouse, 2011).

However, one growing concern associated with the use of MTurk for scholarly work is the naivety, or lack thereof, of its participants (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015). Some MTurk participants, it has been claimed, have become "professional survey-takers,"[1] completing common experimental tasks and questionnaires, often utilized in behavioral research studies, on a daily basis, sometimes more than once. While MTurk does not specifically target the research community, and while there are a variety of tasks (or HITs, for Human Intelligence Tasks) that MTurk workers undertake that are not associated with research, many research studies sample participants from this platform, consequently affecting the level of naivety of the platform. Furthermore, MTurk workers who have completed research tasks for a certain Requester and had a positive experience (in terms of adequacy and timeliness in payments, as well as types of tasks) may be more likely to complete other studies launched by the same Requester, or even similar studies based on the task description, thus reducing the platform's overall level of naivety. The high rate of non-naivety among MTurk participants has recently been shown to have the potential to significantly reduce the effect sizes of known research findings (Chandler et al., 2015). Exacerbating this issue, recent studies have shown that a typical research lab actually samples from an effective population size of only around 7000 participants (and not 500 K, as MTurk advertises), because a small number of MTurk workers are highly active, and consequently usually complete most HITs before other, less active workers have had a chance to see them (Stewart et al., 2015).

Recently, several alternative platforms have emerged, offering services similar to MTurk that could be used for online behavioral research. These alternative platforms offer access to new, more naïve populations than MTurk's, and have fewer restrictions on the types of assignments researchers may ask participants to undertake (Vakharia & Lease,

* Corresponding author.
E-mail address: eyal.peer@biu.ac.il (E. Peer).

[1] See http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory/.

http://dx.doi.org/10.1016/j.jesp.2017.01.006
0022-1031/

2015; Woods, Velasco, Levitan, Wan, & Spence, 2015). For example, MTurk's terms of service prohibit tasks that ask participants to download or install software or applications, or to disclose identifiable personal information (including email addresses). On the other hand, CrowdFlower (CF) – an alternative service – allows for such information to be requested, and imposes the responsibility of due care for confidential data on the requester.[2] Access to alternative crowdsourcing platforms for recruiting human subjects with more naïve populations and fewer limitations could be highly beneficial for researchers interested in conducting online surveys and experiments, as long as these new platforms provide high-quality data.

After searching for and testing several available crowdsourcing websites, we identified and focused on two platforms, similar to Mechanical Turk in design and purpose: CrowdFlower (CF) and Prolific Academic (ProA).[3] CF (https://www.crowdflower.com) was founded in 2007 and is run by executives and a board of directors. This platform is geared towards companies, and boasts a large customer base (including eBay, Microsoft, Cisco, and so on). Some of the use cases listed on CF's website include tasks for sentiment analysis, search relevance, content moderation, data categorization and transcription. CF draws its workforce from a number of different channel partners (such as ClixSense, InstaGC, Persona.ly, and so on), and claims that its workforce includes a broad range of demographics.

ProA (http://www.prolific.ac) was launched in 2014, by a group of graduate students from Oxford and Sheffield Universities, as a software incubator company. It is supported by Isis Innovation, part of the University of Oxford, and is primarily geared towards researchers and startups. ProA provides a range of demographic detail about its participant pool on its website, which researchers can also use to screen participants, suggesting that about 60% of its participants are male, over 70% are Caucasian, and about 50% are students. Table 1 summarizes some key properties and features between these three platforms.

In two studies, we evaluated the data quality of these platforms. In the first study of this paper (Study 1), we compared the data quality of MTurk, CF and ProA, and, as a comparison group, participants from the Center for Behavioral Decision Research (CBDR) participant pool (a more traditional participant pool that includes student and non-student participants, managed by Carnegie Mellon University). Many research institutions have access to participant pools of their own. While they may differ from the CBDR pool, there may also be many commonalities, including composition and retribution models. There is, therefore, much one can learn from by sampling from such a pool and comparing it to participants from online crowdsourcing platforms. In the second study (Study 2), we focused on MTurk and ProA, corroborating the findings from the first study but also expanding the set of tasks used to collect data. In both studies, we compare services along several critical dimensions of online behavioral research. All measures, manipulations, and exclusions in the study are disclosed, as well as the method of determining the final sample size. The authors declare no competing interests. The data and materials for all the studies have been published on the Open Science Framework at https://osf.io/murdt.

1. Study 1

1.1. Method

1.1.1. Sampling and participants

Study 1 consisted of an online survey distributed on four platforms: CF, ProA, CBDR, and MTurk. Our target was to sample about 200 participants from each platform. We limited recruitment time to one week, in order to set a common timeframe for the study. During that week, we were able to reach the goal of recruiting at least 200 participants from each platforms, ending up with a total sample of 831 participants. Table 2 shows the sample size obtained from each platform, the percentage of participants who started but did not complete the study, and the distribution of gender and age in each sample. We conducted the survey on all platforms in January 2016; surveys were submitted on a Thursday during the morning hours (EST); we did not set any restrictions (such as location or previous approval ratings) on any of the platforms, because we wanted to assess differences between the platforms on those aspects too. Participants on MTurk and CF were paid $1 for survey completion; participants on ProA received £1 (equal to $1.47 at the day of the study; payments could only be made in the local currency, and £1 was equivalent to $1 in terms of its proportion of the minimal wage recommended as payment to participants on these sites). Participants on CBDR were given the chance to win a $50 gift card, awarded to one out of every 50 participants. While the expected value of the payment was $1, as in the first two platforms, pilots and previous experience with CBDR samples suggested that the chance of winning a larger prize provides a higher motivation for participation than a certain small payment of $1. Furthermore, the CBDR pool does not offer an online mechanism for compensating participants: they either receive course credit points (if they are students), or are given a monetary reward, such as participation in a lottery.

We found statistically significant differences between the samples in ethnicity, $\chi^2\,(15) = 92.64$, $p < 0.01$, education, $\chi^2\,(6) = 17.85$, $p < 0.01$, and income, $\chi^2\,(18) = 61.5$, $p < 0.01$ (see Appendix for full details). In general, Caucasians were more prevalent on MTurk and ProA than on CF, which included a higher proportion of Asian and Latin/Hispanic participants[4]; CF participants were more educated than the other samples; and MTurk participants had a higher income than the other samples. Regarding location, while the vast majority of MTurk (and CBDR) participants reported[5] that they currently resided in North America (U.S. and Canada), CF and ProA showed a much more diverse distribution across the globe. Not surprisingly, given its location, many ProA participants were from the U.K. and Europe (56% combined), with only 30% from North America, and small percentages from East Asia (4%), Africa (5%) and South America (4%). In CF, in contrast, only 5% came from North America, with the majority of participants from Europe (43%), and another 25% of participants from East Asia or India. The vast majority of participants on MTurk, ProA, and CBDR reported that they could read English at a "very good" or "excellent" level (99%, 97.2%, 91.8%, respectively), versus only 69.2% among CF participants (the rest rated their reading ability as "good" or worse).

1.1.2. Procedure

The study incorporated several stages. The first stage consisted of several questionnaires and experimental tasks adopted from prominent studies in psychology, which were used to assess data quality (adopted from Klein et al., 2014). The second stage included demographic and usage-related questions, designed to better understand the different populations and their use of the different platforms. The last stage included a die-rolling task, designed to test dishonest behavior.

1.1.3. Materials

To examine reliability of data and individual differences between platforms, we used two common scales: the Need for Cognition scale

[2] The terms of service can be found here: https://www.crowdflower.com/legal/.

[3] In addition to CF and ProA, we also examined MicroWorkers, RapidWorkers, MiniJobz, ClickWorker and ShortTask. These websites did not prove as effective as the ones we have chosen to report on – either in their data quality or response rate or the cost of recruitment – and so we do not discuss them in this paper. The details of that preliminary study can be found at https://osf.io/k2nh3/.

[4] The categories we used to measure ethnicity were based on U.S. demographic labels (i.e., Caucasian, African-American, Asian, Latin/Hispanic, and Other). We used these labels similarly across all platforms for the sake of consistency, but these categories might not be interpreted in the same way when dealing with non-US populations. For instance, a "White" European in Spain might identify as "Hispanic."

[5] We compared participants' reported locations to the location of their IP addresses, and confirmed that about 97% of location reports were compatible with the coordinates of their IP address.