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.

Table 1
Comparison of platforms' properties and features (extracted from the platforms' websites).

	MTurk	CF	ProA
Population size	Over 500 K	Over 10 K	About 60 K
Researchers can screen participants			
a) by previous approval rate	Yes, built-in	No option	Yes, built-in
b) by demographics	By location (or creating custom qualifications)	By location and language only	Yes, built-in
c) for taking part in previous studies	By using qualifications	No option	Yes, built-in
Submissions can be automatically checked and approved	No (can set automatic approval for all submissions after preset time)	Yes, using a code on survey completion	Yes, using a code on survey completion
Monetary bonuses can be given to participants	Yes (individually or using a batch file)	Yes, individually	Yes (individually or using a batch file)

(NFC, Cacioppo, Petty, & Kao, 1984), and the Rosenberg Self-Esteem Scale (RSES, Rosenberg, 1979). We selected these scales because (a) they are reliable and validated scales, and (b) they have previously been used successfully to measure data quality on MTurk (Peer, Vosgerau, & Acquisti, 2014). The NFC and RSES use a response scale from 1 (strongly disagree) to 5 (strongly agree). The order of these scales was randomized between participants.

To examine participants' attention, we used four attention-check questions (ACQs; Peer et al., 2014). The details of these ACQs are given in the Appendix. To examine participants' non-naivety (defined as their level of familiarity with commonly used research materials; Chandler et al., 2015), we asked participants to report, after each questionnaire or experimental task, "Was this the first time you were asked to answer such a question/questionnaire?", with options of "yes," "no," and "not sure."

To examine the reproducibility of known effects, we included four judgment and decision-making tasks. The first task was the Asian Disease framing effect (Tversky & Kahneman, 1981), in which participants were asked to imagine that the United States was preparing for the outbreak of a disease, and to select from two courses of action described in either a positive (lives saved) or negative (lives lost) frame: Program A, under which [200 people would be saved] [400 people would die]; or Program B, under which there was a 1/3 probability that 600 people would be saved [no one would die] and 2/3 probability that no one would be saved [600 people would die]. The second task was based on the Sunk Cost Fallacy (following Oppenheimer, Meyvis, & Davidenko, 2009), in which participants were asked to "Imagine that your favorite football team is playing an important game. You have a ticket to the game that you [have paid handsomely for] [have received for free from a friend]. However, on the day of the game, it happens to be freezing cold. What do you do?" Participants rated their likelihood of attending the game from 1 (Definitely stay at home) to 9 (Definitely go to the game). The third task was based on the Retrospective Gambler's Fallacy (Oppenheimer & Monin, 2009), in which participants were asked to "Imagine that you are in a casino and you happen to pass a man rolling dice. You observe him roll three dice and all three come up 6s [one comes up 3 and two come up 6s]. Based on your imagined scenario, how many times do you think the man had rolled the dice before you walked by?" The fourth task was a conceptual replication of the Quote Attribution question (Lorge & Curtis, 1936) in which participants were given the following quote: "I have sworn to only live free, even if I find bitter the taste of death." The quote was attributed to George Washington in one condition and to Osama Bin Laden in the other condition (both persons have been reported to express this statement); participants were asked to indicate how much, on a 7-point scale, they agreed or disagreed with the quote (as used in Chandler et al., 2015). The order of these tasks, as well as the questions within each task, was randomized between participants, and allocation to conditions was randomized within each of these tasks.

After completing all the tasks, participants answered demographic questions, and questions that pertained to the use of their respective platform and other platforms. The final stage of the study included a die-roll "cheating" task. This task was used to examine whether participants would be willing to misreport their performance for additional reward. Participants were told that the survey software would virtually roll a six-sided die, and that the resulting number would be multiplied by 10 cents to determine their bonus for completing the study. However, participants were also told that, before rolling the die, they had to choose whether the bonus would be determined using the upward-facing number on the die, or the number opposite to it, facing downwards. This choice was to be made in their minds before the roll of the die. Then, the die was rolled (using a randomizer) and participants were asked to report the number shown on the die and whether they picked the upward- or downward-facing side, following which they were told what their bonus would be accordingly. Because numbers on opposite sides of a regular six-sided die sum up to 7 and cheating is undetectable, participants had an incentive to cheat, by declaring that they picked the downward-facing side when the side facing up showed a low number, or conversely, that they picked the upward-facing side when the die roll showed a high number on that side. This task was employed only on the platforms that allowed for post-completion monetary bonuses: MTurk, ProA and CF.

Table 2
Sample sizes, dropout rates, workers' demographics.

Sample	Started the study	Completed	Percentage of dropouts	Percent males	Median age (inter-quartile range)
MTurk	220	201	8.6%	56.7%	32.0 (27–38.5)
CF	238	221	7.1%	73.6%	31.0 (25–38)
ProA	243	214	11.9%	64.5%	27.0 (23–37)
CBDR	215	195	9.3%	29.2%	23.5 (23–37)

1.2. Results

1.2.1. Response rates

As detailed in Table 2, dropout rates were around 10% for all platforms, with no significant differences between the platforms, $\chi^2(3) = 3.43$, $p = 0.33$. All of the subsequent analyses include only participants who completed the entire study. Fig. 1 shows the cumulative frequency (absolute number) of accumulated responses according to the time (in minutes) from the onset of the survey, counted from the start time of the first respondent for each sample until the finish time of the last respondent for each sample (which sometimes exceeded 200, as detailed in Table 2). As can be seen, CF showed the fastest response rates, with 200 responses collected within 44 min, followed by MTurk, where it took 1:48 h to collect 200 responses. On ProA, it took 4:37 h to collect 200 responses, and collection was stopped after a week on CBDR (which had provided 195 responses at that time). The average response rate was best on CF and MTurk (3.85 and 5.62 min required for 10 responses), followed by ProA (12.94 min per 10 responses) and CBDR (about 9 h per 10 responses).

To summarize, CF provided a comparable, or even superior, alternative to MTurk in terms of response rate, while ProA had a somewhat slower response rate overall than the two online platforms, but a faster response rate than the university pool. However, if one considers the

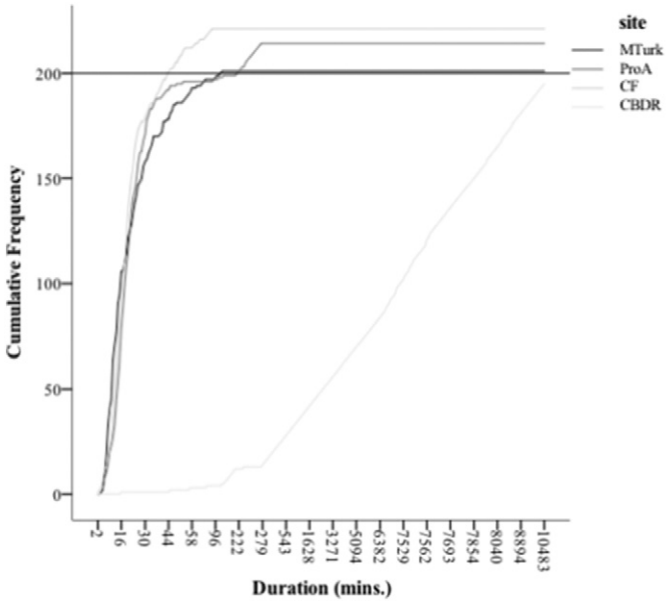

Fig. 1. Response rates across platforms.

time it took each of the three crowdsourcing platforms to reach the 200-responses goal, the difference between ProA and MTurk was less noticeable. We also found some differences in the time taken by participants from the different samples to complete the study. Because the time distribution was highly skewed, we compared medians across groups and found that it was lowest on CBDR (10 min), followed by MTurk (11 min), ProA (14 min), and highest on CF (16 min). A Kruskal-Wallis test showed that these differences were statistically significant ($p < 0.01$).

1.2.2. Attention

Using the four attention-check questions, we tested whether participants read and paid attention to our instructions. In order to capture how researchers might actually use ACQs to exclude inattentive participants, we examined the percent of participant remaining in each sample under two possible exclusion policies: a lenient exclusion policy that excludes all participants that failed more than one ACQ, and a strict exclusion policy that excludes all participants that failed any ACQ. As can be seen in Fig. 2, the strict exclusion policy reduces the sample size by about a half for MTurk, ProA and CBDR, but it is even more detrimental for CF where only 27.1% of participants can be included (χ^2 (3) = 45.19, $p < 0.01$). Using the lenient policy of allowing participants to fail one ACQ reduces the sample size less for all platforms, but still CF's sample is reduced the most to about only 45% of its original size (χ^2 (3) = 80.83, $p < 0.01$).

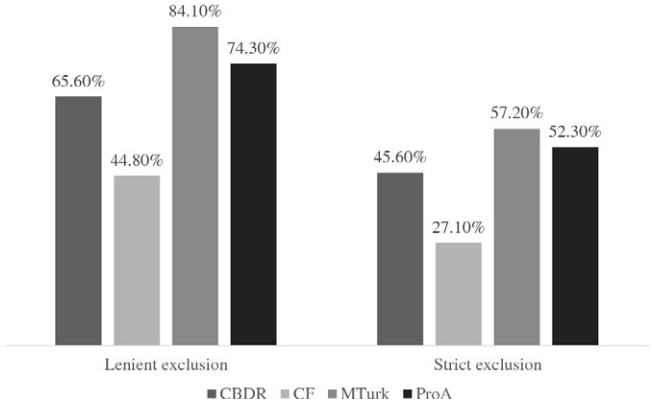

Fig. 2. Percent of participants included in each sample by exclusion policy (lenient = excluding participants that failed more than one ACQ; strict = excluding participants that failed any ACQ).

The average number of failed ACQs also differed significantly between the platforms, F (3, 827) = 37.41, $p < 0.01$. Whereas MTurk participants failed, on average, only 0.67 ACQs (SD = 0.96), ProA participants failed 0.81 ACQs (SD = 1.01); CBDR participants 1.04 ACQs (SD = 1.14); and CF participants failed the most, 1.76 ACQs on average (SD = 1.44). All post-hoc differences, except between ProA and CBDR, were statistically significant after applying Bonferroni's correction ($p < 0.05$). Thus, it appears that CF participants showed the highest, and MTurk participants the lowest, propensity to not follow instructions and fail ACQs; ProA and CBDR participants performed much better than CF, and were only somewhat inferior to MTurk. Because some of the participants from CF reported lower levels of English proficiency, we examined whether this might explain their higher propensity to fail ACQs. We indeed found that CF participants who rated their English proficiency as "good" or less (N = 68, 30.8%) failed, on average, on more ACQs (M = 2.18 vs. 1.58, SD = 1.38, 1.42), t (219) = 2.93, $p < 0.01$. In most cases, failing ACQs probably means that participants did not read the instructions; but it may also suggest that participants' behavior is more naïve and sincere. Thus, to examine this, we included the factor of how many ACQs participants failed in our subsequent analyses of the data quality aspects explored in this study.

1.2.3. Reliability

We compared internal reliability measures (Cronbach's alpha) for the RSES and NFC scales used in the study between platforms, and as a function of exclusion policy. Overall, both scales showed the expected high reliability scores (Cronbach's alpha = 0.898, 0.901 respectively). As shown in Fig. 3, reliability measures for the RSES were adequately high (around or above 0.90) on all platforms except CF, and that did not change considerably under the lenient or strict exclusion policies. For CF, reliability improved significantly (from 0.837 to 0.901) when applying the lenient exclusion policy, and it was similarly high (0.891) under the strict policy. This pattern appeared similarly for the NFC: For all platforms, except CF, reliability was high for the overall sample and also after excluding based on ACQs. For CF, reliability was lower in the overall sample (0.689) and improved significantly (to 0.836) under both exclusion policies. Using Hakstian and Whalen's (1976) method to compare between independent reliability coefficients, we found the differences in reliability among CF, between the groups stated above, were statistically significant for both the RSES and NFC (χ^2 (2) = 8.21, 17.95; $p = 0.02$, $p < 0.01$). We did not find any statistically significant results between the other platforms and their sub-groups.

1.2.4. Reproducibility

We next examined the effect sizes of the four experimental tasks used in the study. We first looked at overall replicability and, as Table 3 shows, found all effects to be statistically significant in MTurk and ProA samples. However, CF participants did not show either the Sunk Cost or Gambler's Fallacy effects. CBDR participants did not exhibit the Gambler's Fallacy effect either. We then examined whether applying an exclusion policy made any difference in any of the platforms. Theoretically, excluding participants that failed ACQs could have two opposing impacts on effect sizes. On one hand, excluding participants reduces sample size and could increase variance that would reduce effect sizes. On the other hand, excluding (presumably) inattentive participants could reduce variance and thus increase effect sizes. Similarly, regarding significance testing, excluding inattentive participants reduces sample size and statistical power while (potentially) reducing variance.

As can be seen in Table 3, excluding participants based on ACQs on MTurk had little to no impact on the observed effect sizes, and it somewhat increased effect sizes on ProA. Among CF participants, the strict exclusion policy had a mixed effect as it increased the effect size of the Asian Disease and Gambler's Fallacy tasks, while it reduced effect sizes

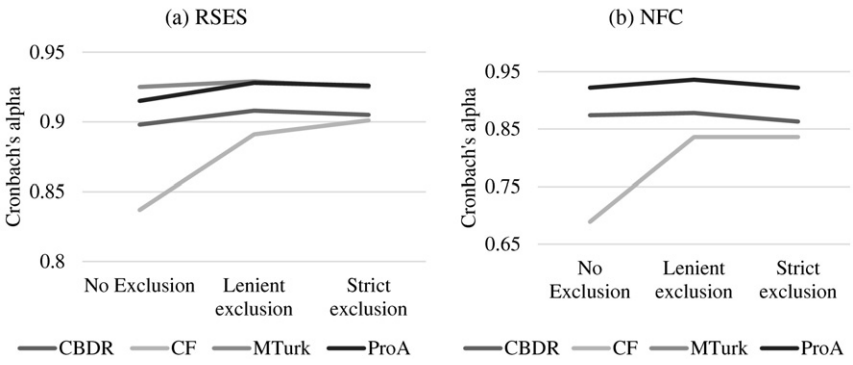

Fig. 3. a–b Cronbach's alpha for the RSES (3a) and NFC (3b) between the platforms and as a function exclusion policy (lenient = excluding participants that failed more than one ACQ; strict = excluding participants that failed any ACQ).

on the other tasks. Among CBDR participants, excluding based on ACQs generally increased effect sizes except for the case of the Quote Attribution task.

1.2.5. Non-naivety

Participants were asked, after each experimental task, questionnaire, and the first two ACQs, whether that was the first time that they had seen that task or question. We coded responses of "yes" as indicating naivety and responses of "no" or "not sure" as indicating familiarity. (Note that "not sure" percentages were <10% across all instances; thus, this classification has little impact on the following results). As Fig. 4 shows, the most familiar tasks were the RSES and NFC scales, followed by the Asian Disease problem. Between the platforms, MTurk participants were typically more familiar with the tasks, while CF participants were more naïve to the tasks.

The reliability of all eight tasks' dichotomous scores of familiarity was adequately high (alpha = 0.744), so we computed the percentage of tasks each participant indicated they were unfamiliar with in order to obtain an overall "naivety" score. ANOVA on the mean percentage of unfamiliar tasks participants reported showed statistically significant differences in naivety between the platforms, F (3, 827) = 25.34, $p < 0.01$. MTurk participants were the least naïve, with an average of 60.3% of tasks reported as seen for the first time, followed by ProA and CBDR (68.3%, 72.2%) participants; CF participants seemed the most naïve, as they reported a mean of 80.8% tasks seen for the first time.

1.2.6. Dishonest behavior

In the last section of the study, participants in all platforms were given the option to cheat by selecting the "up" or "down" side of a randomly rolled die to determine their bonus for completing the study. If all participants were honest, we would expect the mean bonus claimed by participants to be 35 cents (the mean of a uniform distribution of a die roll multiplied by 10 cents). Thus, although we could not determine whether a particular individual participant cheated or not, we could compare the mean bonus claimed in each sample against this benchmark. We found statistically significant degrees of over-reporting in all samples, M = 46.87, 42.29, 40.68, (SD = 12.67, 15.8, 16.18) for MTurk, ProA, and CF participants, respectively, t (200, 213, 220) = 13.27, 6.75, 5.22, $p < 0.01$. However, the effect sizes of cheating degree were significantly highest on MTurk, followed by ProA, and lowest among CF participants, Cohen's d = 1.88, 0.92, 0.70, respectively, F (2, 633) = 9.49, $p < 0.01$. Post-hoc comparisons, using Bonferroni's correction, showed that MTurk's cheating rate was significantly higher than both ProA's and CF's ($p < 0.01$), but that the difference between the latter two samples was not ($p = 0.79$).

1.2.7. Overlap of participants between platforms

We asked participants the frequency with which they used each of the platforms (excluding CBDR, which is not popular among participants worldwide), from "never" to "many times." Table 4 shows the percentage of participants from each platform who reported using other platforms more than "a few times." Generally, the degree of overlap between platforms seems to be quite small, with the highest overlap among the 22% of ProA users who also used MTurk.

1.2.8. Usage patterns

As can be seen in Fig.s 5, 77.2% of MTurk, and 84.2% of CF participants reported spending 8 or more hours per week on the platform. ProA users spent considerably less time, with 69.1% reporting spending between 1 and 8 h per week. As Fig. 6 shows, this difference in usage clearly resulted in earning differences between the platforms: whereas >70% of MTurk-ers reported earning more than $50 a week, about 72% of CF participants reported earning $5–$50 a week, and 77% of ProA participants reported earning less than $10 a week (76% of CBDR participants reported earning less than $5 a week, possibly due to students receiving academic credit instead of money). The differences between average

Table 3
Effect sizes (Cohen's d) between platforms and exclusion policies.

Platform	Exclusion policy	Asian disease	Sunk cost	Gambler's Fallacy[a]	Quote attribution
MTurk	None (all Ps.)	0.82	0.27	0.28	0.73
	Lenient exclusion	0.99	0.34	0.29	0.75
	Strict exclusion	0.94	0.24	*0.24*	0.73
ProA	None (all Ps.)	0.63	0.39	0.29	0.68
	Lenient exclusion	0.74	0.61	0.36	0.66
	Strict exclusion	0.82	0.53	*0.31*	0.72
CF	None (all Ps.)	0.72	*0.02*	*0.20*	0.54
	Lenient exclusion	0.82	*−0.29*	0.39	0.38
	Strict exclusion	0.76	*−0.62*	*0.35*	*0.25*
CBDR	None (all Ps.)	0.76	0.42	*0.12*	0.51
	Lenient exclusion	1.11	0.41	*0.14*	0.28
	Strict exclusion	1.12	0.56	*0.25*	*−0.01*

Note: all effect sizes were statistically significant, $p < 0.05$, except for those that are in italics.
[a] We excluded responses of above 100, which constituted <5% of the data.

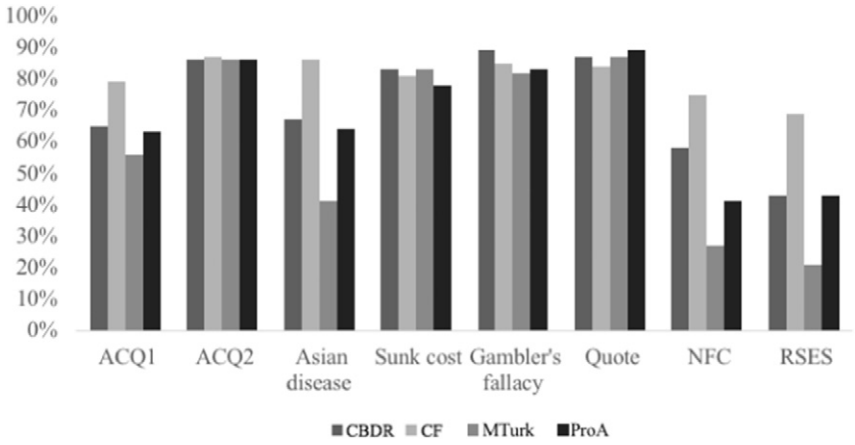

Fig. 4. Percentage of naïve participants (not familiar with the task) per task per platform.

pay/week between the samples were statistically significant, F (3, 769) = 371.46, $p < 0.01$, as MTurk participants reported the highest average pay (M = $3.69, SD = $0.5), followed by CF (M = $2.54, SD = $0.9), ProA (M = $1.81, SD = $0.81) and CBDR (M = $1.3, SD = $0.57). All the pairwise differences between the platforms were statistically significant, $p < 0.01$, after Bonferroni's correction. Consistently, the median number of tasks participants reported completing on the platform was highest among MTurk (7100), lower on CF (1000) and much lower on ProA (30) and CBDR (6). The median approval score (percentage of approved submissions) participants reported having was close to 100% for all platforms except for CF (89%).

1.3. Discussion

To summarize the comparison of data quality between the platforms, we found that, compared to MTurk, CF participants showed a higher response rate but also a much higher rate of failing attention-check questions, resulting in lower values of internal reliability for the participants on CF who failed ACQs. Additionally, while CF participants reported less familiarity (higher naivety) regarding common experimental tasks, the effects for two of these tasks could not be replicated on that sample, whereas effects for all tasks replicated on ProA. In addition, ProA participants reported higher naivety than MTurk participants. Lastly, both ProA and CF participants showed lower degrees of dishonest behavior, compared to MTurk. A summary comparison of the differences found between the platforms is given in Table 5.

These results suggest that while both CF and ProA show adequate data quality, ProA seems to be the most viable alternative to MTurk. ProA users showed only slightly lower levels of attention as compared to MTurk, which did not significantly affect measures of reliability. Furthermore, with a higher level of naivety and lower frequencies of weekly participation as compared to MTurk, the ProA sample reproduced known effects of all the tested tasks, while only half were reproduced on CF. Finally, we observed a lower propensity on the part of ProA participants to engage in dishonest behavior, as compared to MTurk. Overall, ProA demonstrated superiority over CF. However, it took longer to collect all responses, and data collection on ProA slowed down significantly as we approached the 200-participant mark (for the first 180 participants, ProA proved to be the fastest route to collect data). This might be a symptom of the smaller overall size of ProA, as compared to CF (and MTurk). ProA users also scored significantly higher on the attention checks as compared to CF. The higher rates of passing attention-check questions on ProA (and MTurk) could be due to participants' past experience with these or similar attention-check questions (Chandler, Mueller, & Paolacci, 2014; Peer et al., 2014), and a high failure rate could actually be considered desirable because it implies naivety with regards to experimental materials. Notwithstanding higher naivety, one should consider the failure in replicating both the Sunk Cost and the Gambler's Fallacy effects on CF, which may be especially worrisome for the psychology research community.

Propensity to cheat, on the other hand, was not statistically different between CF and ProA: participants on both of these platforms exhibited a lower propensity towards cheating, as compared to MTurk. This could be due to a number of reasons, including (but not limited to): the specific task or incentive scheme we used; participants' familiarity with the task; participants' suspicion that they might be monitored; or participants' general reluctance to expose their true behavioral tendencies. Alternatively, this could be due to individual differences between the participants in the different samples, or also related to the platform itself: while ProA advertises itself as designed for academic research, MTurk's appeal is more about earning money quickly.

When researchers choose between platforms, they should consider two other issues raised by our data. First, although we found no substantial overlap between participants from CF and MTurk (<10% of participants reported using both platforms), some participants (about 22%) from ProA indicated that they use MTurk as well. This should not be an issue if one restricts the study to a single platform, but should be taken into account if the study is to be run on multiple platforms, or if (for instance) a similar study has already been conducted on one of the platforms. The other issue to consider is the demographic composition of these platforms. The most salient difference lies in participants' ethnicity and country of origin. Whereas CF participants showed the highest diversity in terms of ethnicity, ProA's distribution was similar to MTurk's, with a lower percentage of non-Caucasian participants. Moreover, a large portion of CF and ProA participants reside outside the U.S. (mainly in Europe and Asia), while MTurk attracts mostly U.S. residents. This suggests that the different platforms tap into different populations, and this should be taken into account when determining which platform to use for participant recruitment.

These differences in demographic and geographic origin between the platforms, and especially between CF and MTurk, deserve special attention. On one hand, the differences in both ethnicity and country of residence between these two platforms suggest that one is not comparable with the other, and thus CF cannot be considered a comparable

Table 4
Percentage of participants reporting using platforms more than "a few times."

	Uses MTurk	Uses CF	Uses ProA
MTurk	98.50%	2.5%	14.5%
CF	6.3%	94.1%	4.1%
ProA	22%	9.3%	88.8%
CBDR	8.3%	1.5%	1%

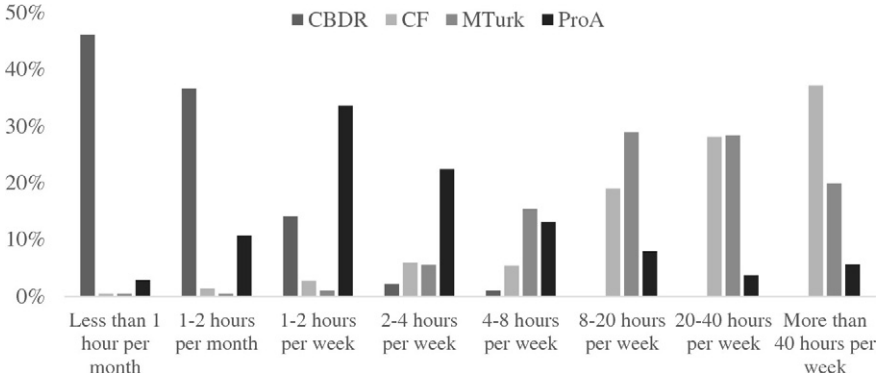

Fig. 5. Distribution of frequency of usage between the platforms.

alternative to MTurk. On the other hand, scholars have urged the scientific community to expand beyond western, industrialized, educated, rich and democratic participants (or WEIRD; see Henrich, Heine, & Norenzayan, 2010), and specifically beyond U.S.-based participants, which, as our results suggest, are over-represented on MTurk. In that sense, researchers may choose to take advantage of CF's or ProA's access to non-U.S. populations. In doing so, researchers may also benefit from this population's relative naivety towards many behavioral and psychological research materials, a point that has been singled out as one of MTurk's most persistent disadvantages (Chandler et al., 2014).

Overall, the results of our first study suggest that ProA (but not CF) could be considered a potential alternative to MTurk as it produced data quality of comparable levels, with more diverse and naïve participants, at a reasonable (albeit slower) response rate. However, while many studies have examined MTurk's data quality (as reviewed in Paolacci & Chandler, 2014), the study above constitutes the first systematic examination of ProA's data quality.

Despite their value, though, we cannot and probably should not treat these findings as final. Additionally, after Study 1 was conducted, ProA changed their pricing scheme to significantly raise the commission paid by researchers. It thus seemed pertinent to re-evaluate ProA as some dimensions (e.g., response rates) may have been affected by that change. In order to verify that ProA may be considered as an alternative to MTurk, we conducted a second study, in which we focused on ProA and MTurk alone, and with a much larger sample.

2. Study 2

2.1. Method

2.1.1. Samples' composition and characteristics

We recruited 1374 participants from both sites (691 from MTurk and 683 from ProA), of which 1205 (604 from MTurk and 601 from ProA) completed the entire survey. Because Study 2 occurred a year after Study 1 was completed, and because tasks differed across the two studies, we did not screen out participants that completed Study 1. Participants were paid $1 on MTurk and £1 on ProA (equal to $1.23 at the day of the study). Dropout rates were similar for MTurk and ProA (12.6% and 12.0%, respectively). From here on, we analyzed only the results of those who completed the entire study. There were no differences in gender between the sites (53.1% vs. 56.1% males on MTurk vs. ProA, χ^2 (1) = 1.04, $p = 0.31$) but MTurk participants were somewhat older than ProA's (Md_{age} = 32 vs. 28.5, inter-quartile range = 28–42, 24–35, respectively, $p < 0.01$). We found statistically significant differences between the sites in ethnicity, χ^2 (5) = 25.51, $p < 0.01$, education, χ^2 (9) = 60.04, $p < 0.01$, and income, χ^2 (7) = 147.02, $p < 0.01$, but not in English proficiency, t (0.05), $p = 0.96$ (see Appendix for more details). In general, ProA participants included slightly more Asians and Hispanics, and slightly fewer African-American and Caucasians than MTurk; and they were somewhat more educated and had lower income compared to MTurk. The reported location[6] of participants differed significantly between the sites, χ^2 (6) = 575.2, $p < 0.01$. While 90.5% of MTurk participants were from North America, and 6.8% from India (the rest came from Europe, East Asia, Africa and the U.K.), North Americans comprised only 25.9% of ProA's participants, which also included 30.8% from the U.K., another 27.1% from Europe, 8.1% from South America, and 6% from India (the rest were from Africa and East Asia).

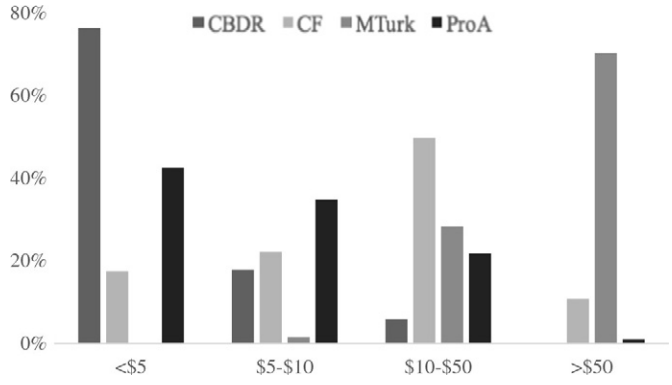

Fig. 6. Percentage of participants in different quartiles of average weekly earning between the platforms (the cutoffs represent the quartiles of earnings in the overall sample).

2.1.2. Procedure

Participants were invited to complete an online study that consisted of the following parts. To assess reliability, we used the Consideration for Future Consequences scale (Strathman, Gleicher, Boninger, & Edwards, 1994). To examine attention, we included three ACQs. One was an item embedded into the CFC scale ("I think I have never used the Internet myself at any time through the course of my personal life" – any answer other than "1-extremely uncharacteristic" was coded as failing the ACQ). Another ACQ was a fake "perceptual abilities task." We told participants that they would see an image with many people in it and that their task was to count how many persons appear in the picture within 10 s. However, in the text describing the task we instructed participants to actually report zero. The third ACQ was a short questionnaire about liking math, that had three items. In the introduction to the questionnaire, we asked participants to answer "six" for the first item, to divide that number by two and use the result as the answer for the second and third questions. To examine reproducibility of known effects we used the "simulation heuristic" (Kahneman & Tversky, 1982), in which participants read that "Mrs. Crane and Mrs. Tees were scheduled to leave the airport at the same time, but on different flights. Each of them woke up and left

[6] Participants' reported locations matched their IP addresses in 94% of the cases.

Table 5
Summary of differences between the platforms.

	MTurk	CF	ProA	CBDR
Dropout rate	Low	Low	Low	Low
Response rate	Fast	Fastest	Fast	Slowest
ACQs failure rate	Lowest	Highest	Low	Medium
Reliability	High	Low	High	High
Reproducibility	Good	Poor	Good	Fair
Naivety	Lowest	Highest	High	High
Dishonesty	Highest	Medium	Medium	–
Ethnic diversity	Low	High	Low	Medium
Geographic origin	Mostly U.S.	Mainly Europe	Mostly U.S.	Mostly U.S.
English fluency	High	Low	High	High
Income level	Low	Low	Medium	Low
Median education level	Bachelor's	Bachelor's	Bachelor's	Bachelor's
Usage frequency	High	Highest	Medium	Lowest
Overlap with other	Some (ProA)	Few	Some (MTurk)	Few

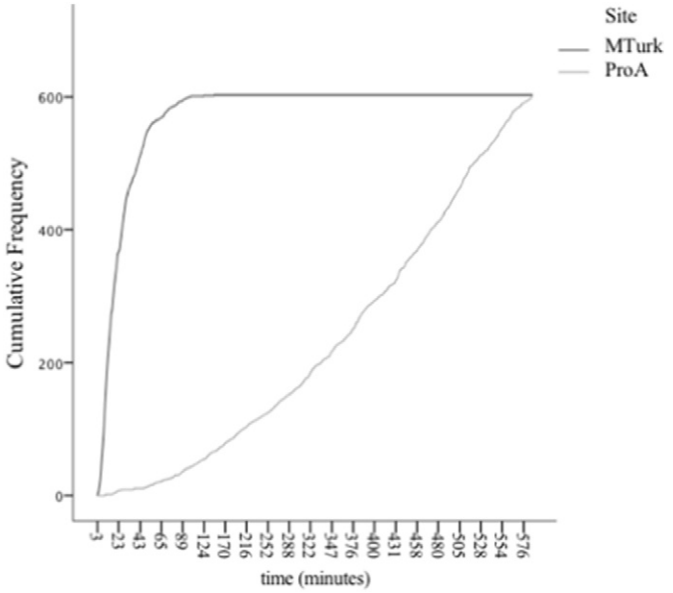

Fig. 7. Response rates between the two platforms.

home at the same time, drove the same distance to the airport, was caught in a traffic jam, and arrived at the airport 30 minutes after the scheduled departure of their flights. Mrs. Crane was told at her gate that her flight left on time. Mrs. Tees was told at her gate that her flight was delayed and had left just three minutes ago. They both had dawdled for ten minutes before leaving home." Participants were asked to indicate who, between Mrs. Crane and Mrs. Tees, they felt her dawdling to be more foolish (or irresponsible). Responses were entered on a 7-point scale ranging from "Mrs. Tees felt more foolish considerably" to "Mrs. Crane felt more foolish considerably." Typically, respondents think the person who missed the flight by a short duration should feel more regret, thus exhibiting counterfactual thinking.

Participants then completed demographic and usage-related questions similar to Study 1. We also included questions that were designed to test some other hypotheses (for example, we asked participants how many shoes they owned in order to test the hypothesis that women have more shoes than men). The purpose of these questions was to allow us to examine the effect sizes of such "obvious" hypotheses that could then be used to calculate the minimum sample size required, on each platform, to obtain a statistically significant result for that hypothesis. However, these results ended up being ambiguous in interpretation and, under editorial advice, we decided to exclude them from the paper. Interested readers may find the full details of these questions and results at https://osf.io/7ut8h. All of the above parts were given to participants in random order.

2.2. Results

2.2.1. Response rates

As Fig. 7 shows, the response rate on MTurk was much faster than on ProA in this study. While data collection was completed on MTurk in 151 min, it took almost 10 h to reach 600 responses on ProA. This means the response rate was almost four times faster on MTurk (3.99 responses per minute on MTurk vs. 1.01 responses per minute on ProA).

2.2.2. Attention

While 60.6% of participants on MTurk passed all three ACQs, only 48.4% of ProA's participants passed all ACQs. The percent of participants failing one, two or all ACQs on MTurk were 26%, 9.6% and 3.8% while on ProA these were 19%, 20% and 12%. These differences, which were statistically significant, χ^2 (3) = 64.03, $p < 0.01$, suggest a higher overall failure rate of ACQs on ProA compared to MTurk. Respectively, we found that a lenient exclusion policy, excluding participants who failed more than one ACQ, would result in retaining 86.6% of the sample on MTurk compared to only 67.6% on ProA, χ^2 (1) = 61.18, $p < 0.01$

2.2.3. Reliability

After coding the CFC questions according to the scale, we examined its reliability between the sites and between exclusion policies. As can be seen in Fig. 8, reliability on MTurk was found to be slightly higher than on ProA when using all participants (Cronbach's alpha = 0.89 vs. 0.821). This difference was slightly minimized under the lenient exclusion policy (0.911 vs. 0.869) and the strict policy (0.904 vs. 0.863). Using Hakstian and Whalen's (1976) method, we found that the differences between all reliability coefficients were statistically significant (χ^2 (7) = 153.58, $p < 0.001$). However, it should be noted that in all instances reliability remained above the conventional threshold of 0.8 indicating adequate reliability.

2.2.4. Reproducibility

The simulation heuristic predicts that people would believe that a person who missed their flight by a few minutes would feel more regret (i.e., feel more foolish about dawdling before leaving for the airport) than a person who missed their flight by a longer duration. As Fig. 9 shows, the effect, which is indicated by a mean regret rating that is significantly higher than the scale's midpoint (4), was found on both sites. The effect was slightly stronger under the exclusion policies, but these

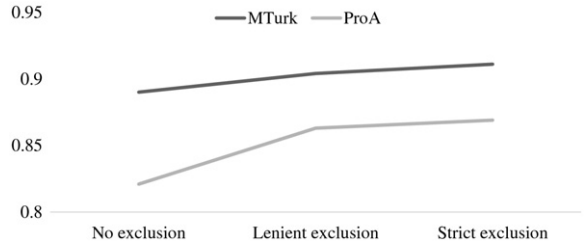

Fig. 8. Cronbach's alpha for the CFC scale as a function of platforms and exclusion policy (lenient = excluding participants that failed more than one ACQ, strict = excluding participants that failed any ACQ).

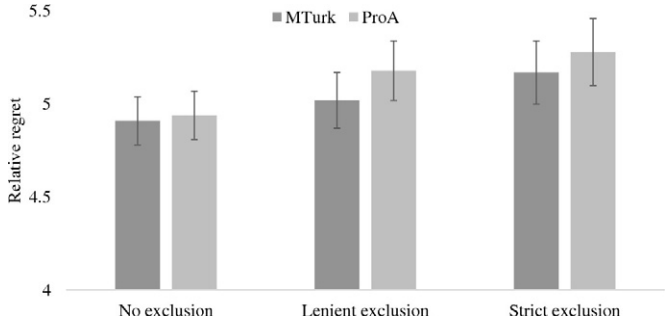

Fig. 9. Relative regret ratings as a function of platforms and exclusion policy (higher scores indicate greater expected regret on the part of the person who missed the flight by a little).

differences were not statistically significant, as is evident by the overlap of the 95% confidence interval bars in Fig. 9.

To summarize thus far, it appears that ProA had a significantly lower response rate, and ProA participants failed ACQs somewhat more often than MTurk participants. Reliability was high on both sites, with MTurk showing somewhat higher reliability. Excluding participants based on ACQs improved reliability on both sites. On both sites, the simulation heuristic was replicated successfully, with no significant differences between the sites. Thus, it appears that both sites provide high data quality on all the examined parameters.

2.2.5. Usage patterns

As can be seen in Fig. 10, most MTurk participants reported spending between 8 and 20 h per week on the platform. ProA users spent considerably less time, most reporting spending between 1 and 2 h per week only. As Fig. 11 shows, this clearly results in earning differences between the platforms: whereas > 70% of participants on MTurk reported earning more than $50 a week, about 85% of ProA participants reported earning less than $10 a week. Consistently, the median of the total number of tasks participants reported completing in their lifetime as a participant on that platform was much higher on MTurk (5900), than ProA (10). This is consistent with the fact that MTurk has been available for several years before ProA was launched. The median approval score (percentage of approved submissions) participants reported was close to 100% for both platforms.

3. General discussion

Some of the results of Study 2 corroborated the findings of Study 1, while others were different. Similar to Study 1, we found that both MTurk and ProA produced high-quality data for many of the aspects examined in the study. The rate of attention was quite high on both platforms, with a majority of participants passing all ACQs (or failing only one). Again, MTurk participants showed higher rates of passing ACQs compared to ProA. Reliability remained high on both platforms, and it

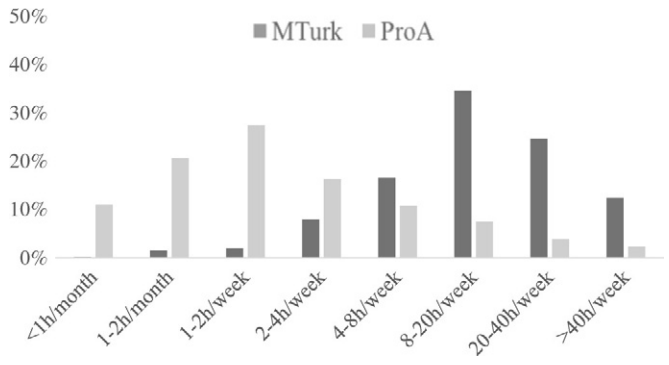

Fig. 10. Distribution of frequency of usage between the platforms.

Fig. 11. Distribution of average weekly earning between the platforms.

remained consistently high when excluding participants who failed ACQs, on both sites. The results suggest that on both MTurk and ProA, most of the participants pay attention to instructions and consistently complete questionnaires carefully. We were also able to replicate the simulation heuristic on both platforms, also when excluding participants based on ACQs. This shows that both sites' participants provide high data quality, even when some fail some of the ACQs. This ceiling effect of ACQs on data quality is consistent with Peer et al. (2014) which showed that ACQs have low diagnostic ability when data quality is already high.

In contrast to Study 1, response rates between MTurk and ProA were considerably different. Whereas in Study 1 the difference between the response rate on MTurk to ProA was about 2.5 times in favor of MTurk, that ratio increased to 4 times in favor of MTurk in Study 2. This could be due to the fact that we sampled three times more participants in this study, and also because of the fact that in the period between the studies, ProA changed its pricing scheme. The change in pricing scheme, which significantly raised commissions for researchers (from a 10% flat rate commission to 12.5% + 10 p per participant), might have influenced how researchers, and participants, use the site. For instance, researchers may have begun to run studies in bulk batches, in order to reduce the effective rates of commission they pay. If so, this would result in fewer individual studies posted online, which may have increased the share of lengthy studies offered to participants; it is reasonable to speculate that this might deter some participants from using the site, which would affect the response rate. It is also possible that the actual overall number of active participants on ProA is less than ProA's advertised rate.

To summarize, our studies show that the major advantage of MTurk over ProA lies in its faster response times. While slower than MTurk, ProA provides data quality that is comparable or not significantly different than MTurk's, and ProA's participants seem to be more naïve to common experimental research tasks, and offer a more diverse population in terms of geographical location, ethnicity, etc. This suggests to researchers who are more interested in obtaining results faster, from a more homogeneous sample, that they should use MTurk, while researchers who prefer naivety and diversity in their sample, could turn to ProA if they are willing to wait some more for data collection to complete (depending on sample size).

While the results of the current research can serve to present researchers with a range of choices when venturing with online crowdsourcing research, additional research is necessary to explore some of the unanswered questions emerging from the current studies' limitations. First, the roots and causes of the differences found between the platforms remain unclear, as we could only control the sampling (and not allocation) of participants from the different platforms. Second, it remains an open question how constant or transient any of the findings may be. While some differences seemed to be relatively stable (e.g., demographics), many others (e.g., response rates, naivety, and so forth) could be much more temporary. In this regard, the current paper offers a helpful framework through which platforms can be evaluated over time (and also following certain events, such as a major change in pricing). This framework, which includes measures of attention, reliability, reproducibility, naivety and dishonesty, could also be used to evaluate new platforms that may arise in the future to present researchers with new capabilities for conducting experimental and behavioral research online.

Appendix A – Additional figures

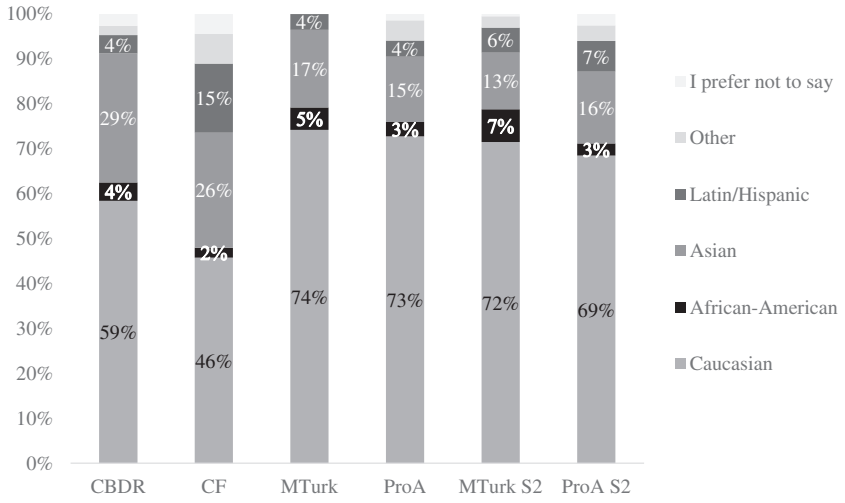

Fig. A1. Ethnicity distributions, from Studies 1 and 2 (S2 refers to Study 2). Note: the same categories and labels were used on all platforms.

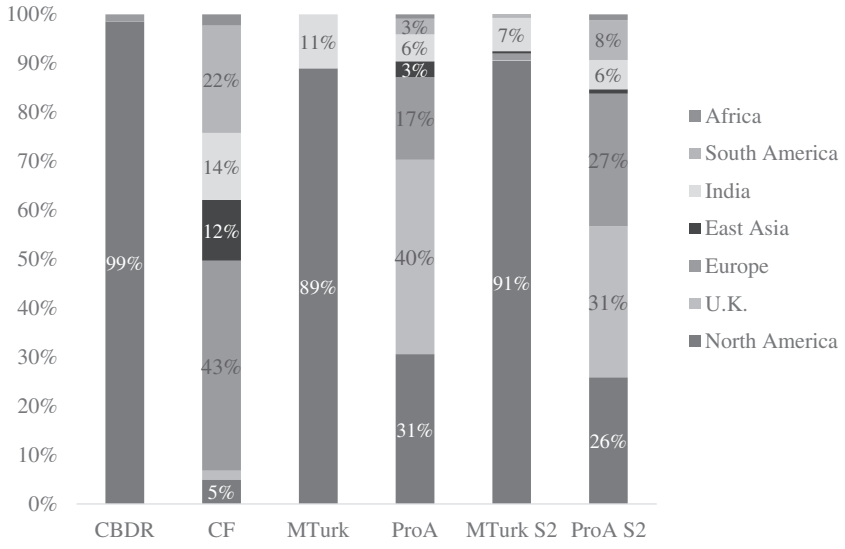

Fig. A2. Reported location distributions from Studies 1 and 2 (S2 refers to Study 2).

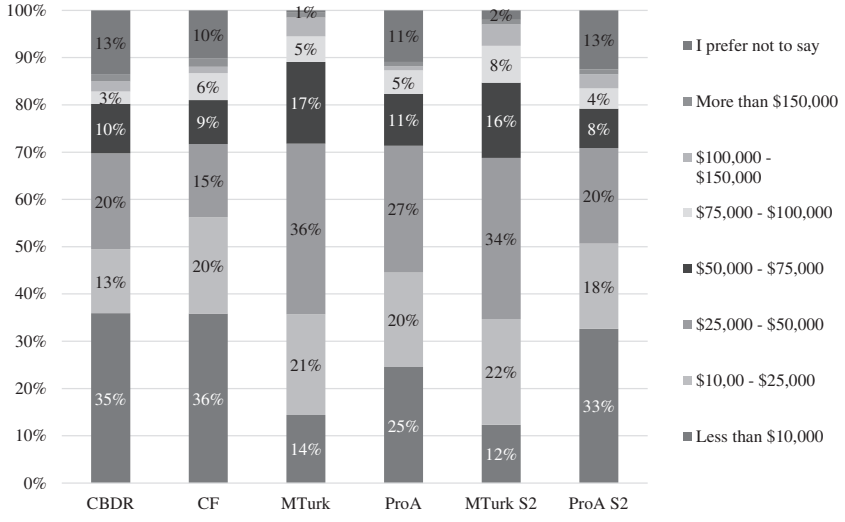

Fig. A3. Reported income distributions from Studies 1 and 2 (S2 refers to Study 2).

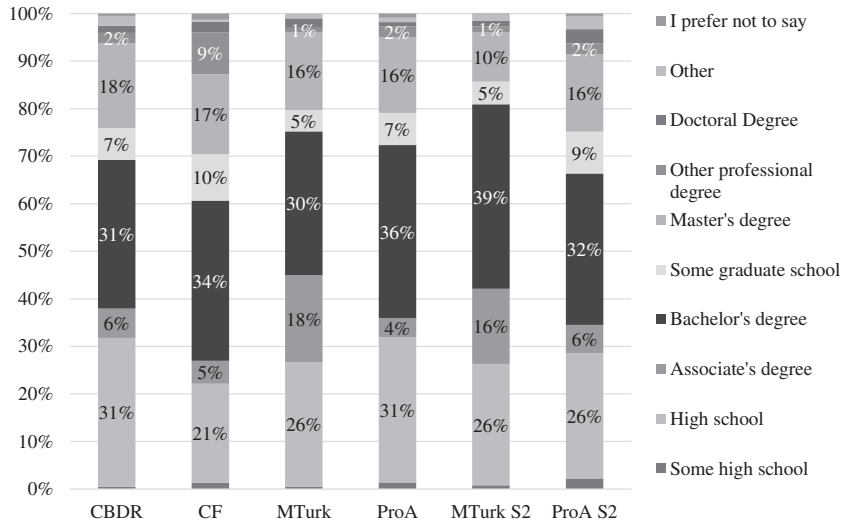

Fig. A4. Reported education level distributions from Studies 1 and 2 (S2 refers to Study 2).

Appendix B. Supplementary data

Supplementary data and full research materials to this article can be found online at https://osf.io/murdt.

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307.

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., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. (2015). Non-naïve participants can reduce effect sizes. *Psychological Science*, *26*(7), 1131–1139.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, *37*(2), 413–420.

Goodman, J. K., Cryder, C. E., & 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.

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219–231.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge Univ. Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š, Bernstein, M. J., & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142–152.

Lorge, I., & Curtis, C. C. (1936). Prestige, suggestion, and attitudes. *The Journal of Social Psychology*, *7*(4), 386–402.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision making*, *4*(5), 326.

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

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, *5*(5), 411–419.

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.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.

Rosenberg, M. (1979). *Rosenberg self-esteem scale*. New York: Basic Books.

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, *46*(1), 95–111.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making*, *10*(5), 479–491.

Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology*, *66*(4), 742–775.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Vakharia, D., & Lease, M. (2015). Beyond Mechanical Turk: An analysis of paid crowd work platforms. *Proceedings of iConference 2015* Retrieved online at April 14, 2015, from https://www.ischool.utexas.edu/~ml/papers/donna-iconf15.pdf

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ*, *3*, e1058.