Beyond the Turk:

Alternative platforms for crowdsourcing behavioral research

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

[a] Corresponding author. Graduate School of Business Administration, Bar-Ilan University, Ramat-Gan, 52900, Israel. eyal.peer@biu.ac.il

[b] Eller College of Management, University of Arizona, Tucson, AZ.

[c] Heinz College, Carnegie Mellon University, Pittsburgh, PA.

Beyond the Turk: Alternative platforms for crowdsourcing behavioral research

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.

Key words: online research; crowdsourcing; data quality; Amazon Mechanical Turk; Prolific Academic; CrowdFlower

Beyond the Turk: Alternative platforms for crowdsourcing behavioral research

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, Muller, & Ratkliff, 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

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

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 7,000 participants (and not 500K, 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, 2015; Woods et al., 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.

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

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

	MTurk	**CF**	**ProA**
Population size	Over 500K	Over 10K	About 60K
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)

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,

[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/

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 these alternative platforms with data collected via both MTurk and a university-based online participant pool. Study 1 included all three online platforms 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 has been published on the Open Science Framework at https://osf.io/murdt.

Study 1

Method

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, χ^2 (15) = 92.64, $p < .01$, education, χ^2 (6) = 17.85, $p < .01$, and income, χ^2 (18) = 61.5, $p < .01$ (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

[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.

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).

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.

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)

Materials. To examine reliability of data and individual differences between platforms, we used two common scales: The Need for Cognition scale (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, 2013). 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., 2013). 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.

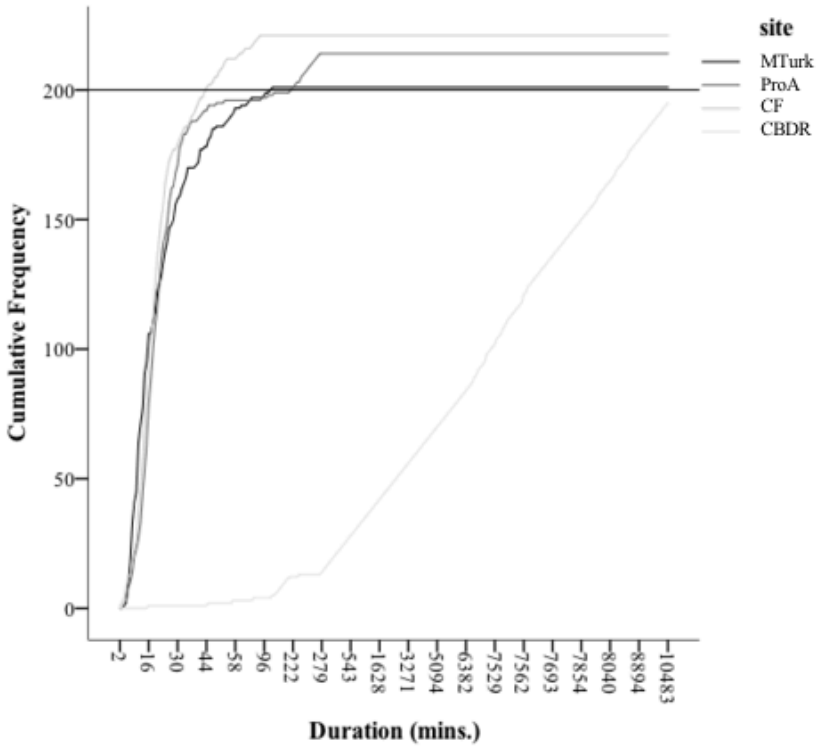
Results

Response rates. As detailed in Table 2, dropout rates were around 10% for all platforms, with no significant differences between the platforms, χ^2 (3) = 3.43, p = .33. All of the subsequent analyses include only participants who completed the entire study. Figure 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 minutes, followed by MTurk, where it took 1:48 hours to collect 200 responses. On ProA, it took 4:37 hours 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 minutes required for 10 responses), followed by ProA (12.94 minutes per 10 responses) and CBDR (about 9 hours 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 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 minutes), followed by MTurk (11 minutes), ProA (14 minutes), and highest on CF (16 minutes). A Kruskal-Wallis test showed that these differences were statistically significant ($p < .01$).

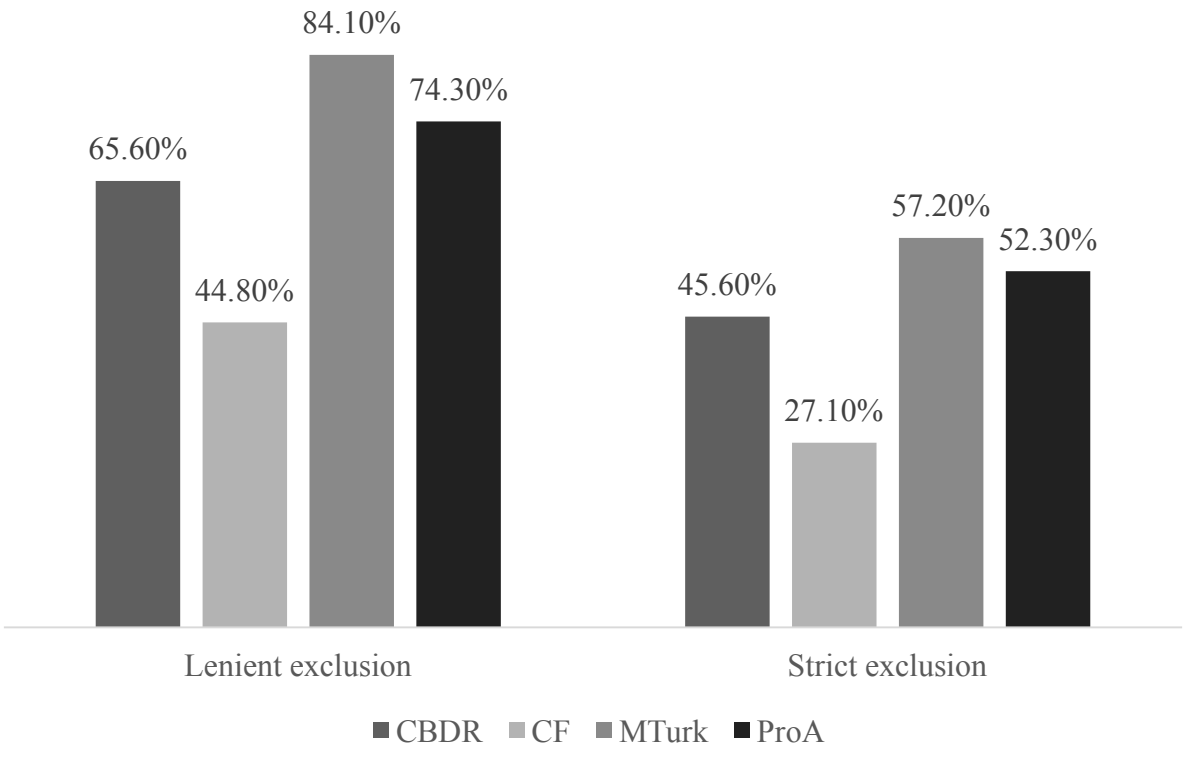
Figure 1. Response rates across platforms.



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 Figure 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, $p < .01$). 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, $p < .01$).

Figure 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).

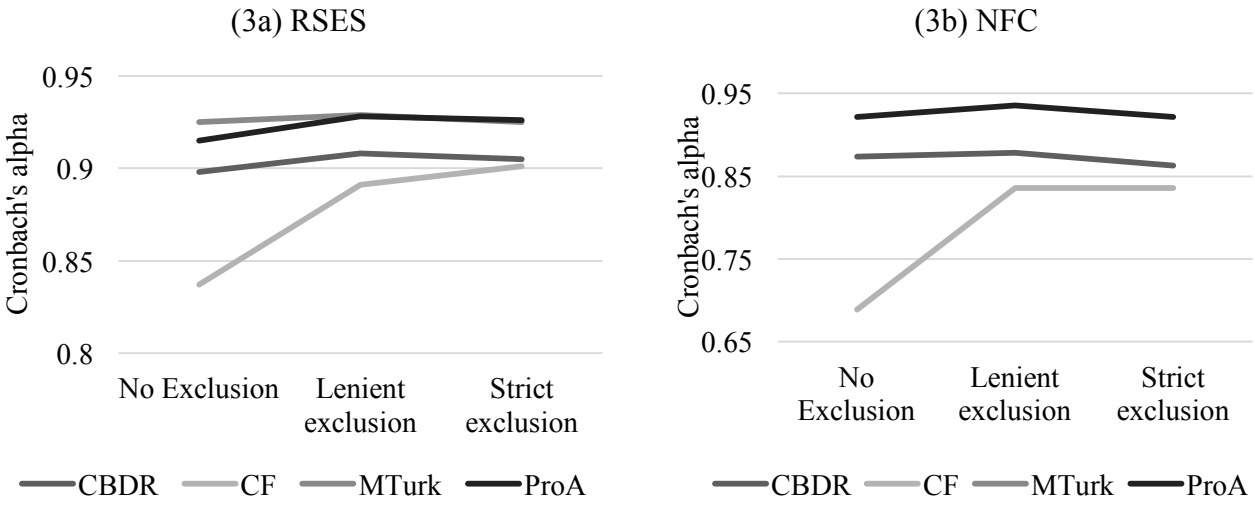


The average number of failed ACQs also differed significantly between the platforms, F (3, 827) = 37.41, $p < .01$. Whereas MTurk participants failed, on average, only 0.67 ACQs (SD=0.96), ProA participants failed 0.81 ACQs (SD=1.01); CBDR participants 1.04 ACQs (SD=1.14); and CF participants failed 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 ($p < .05$). 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, $p < .01$. 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.

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 Figure 3a, 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 = .02, p < .01). We did not find any statistically significant results between the other platforms and their sub-groups.

Figure 3a-3b. 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).



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.

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

Platform	Exclusion policy	Asian Disease	Sunk Cost	Gambler's Fallacy*	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, $p < .05$, except for those that are in italics.

* We excluded responses of above 100, which constituted less than 5% of the data.

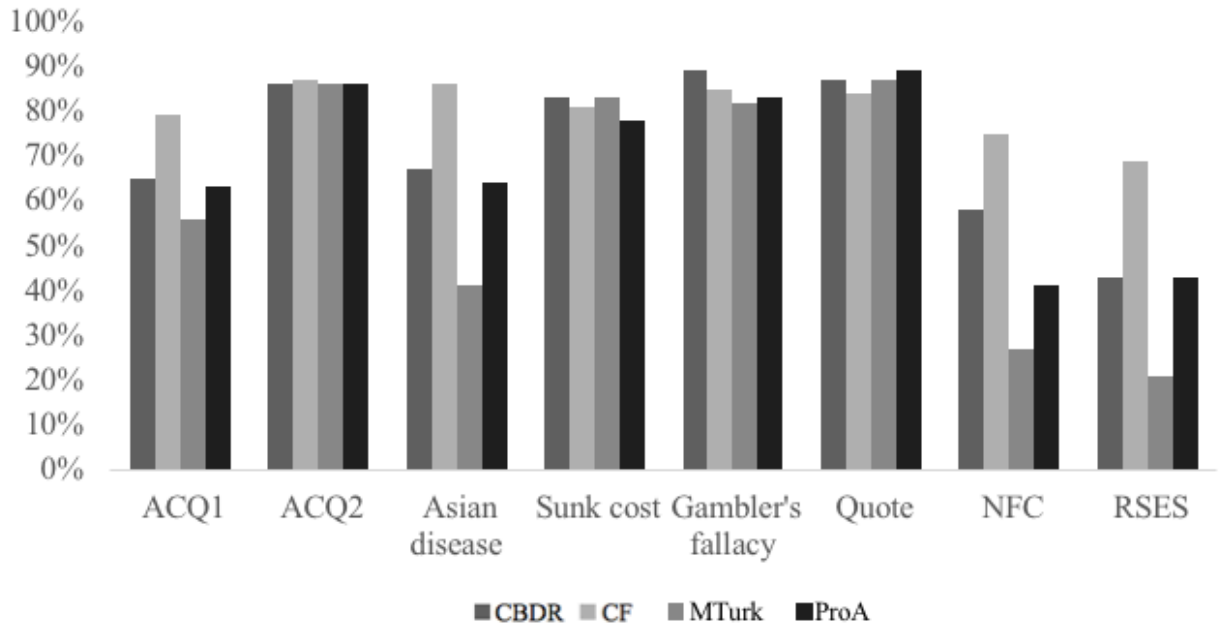
As can be seen in Table 3, excluding participants based on ACQs on MTurk had little to no impact on the observed effect sizes, and it somewhat increased effect sizes on ProA. Among CF participants, the strict exclusion policy had a mixed effect as it increased the effect size of the Asian Disease and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

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 less than 10% across all instances; thus, this classification has little impact on the following results). As Figure 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, $F(3, 827) = 25.34$, $p < .01$. 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.

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



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, $p < .01$. 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, $p < .01$. Post-hoc

comparisons, using Bonferroni's correction, showed that MTurk's cheating rate was significantly higher than both ProA's and CF's ($p < .01$), but that the difference between the latter two samples was not ($p = 0.79$).

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.

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%

Usage patterns. As can be seen in Figure 8, 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 to 8 hours per week. As Figure 9 shows, this difference in usage clearly resulted in earning differences between the platforms: whereas more than 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 pay/week between the samples were statistically significant, $F(3, 769) = 371.46$, $p < .01$, 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, $p < .01$, after Bonferroni's correction. Consistently, the median number of tasks participants reported completing on the platform was highest among MTurk (7,100), lower on CF (1,000) 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%).

Figure 8. Distribution of frequency of usage between the platforms.

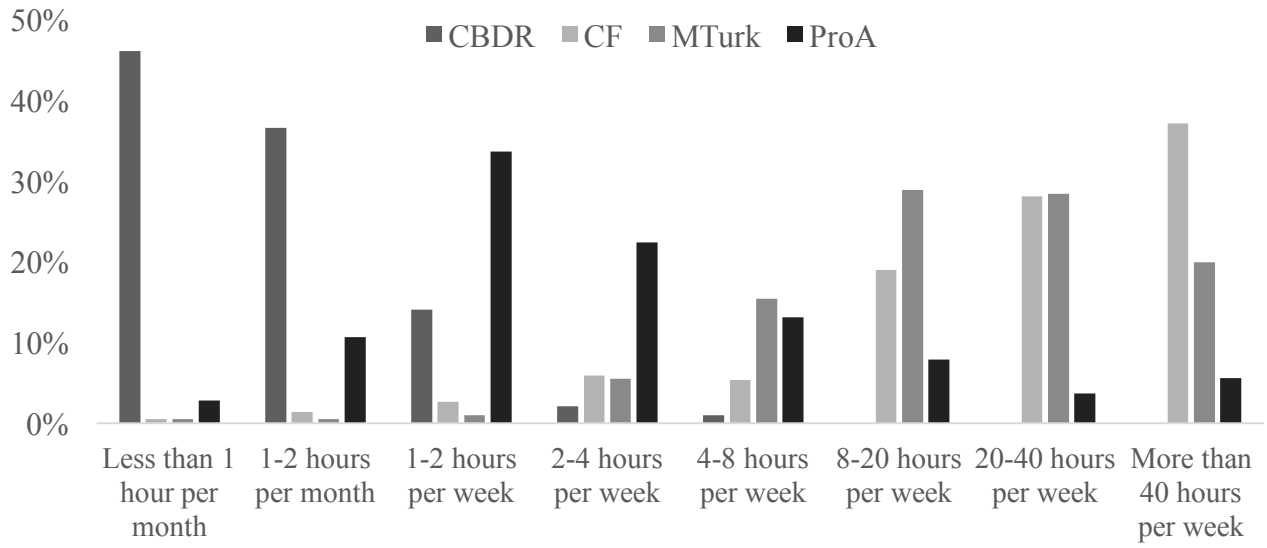
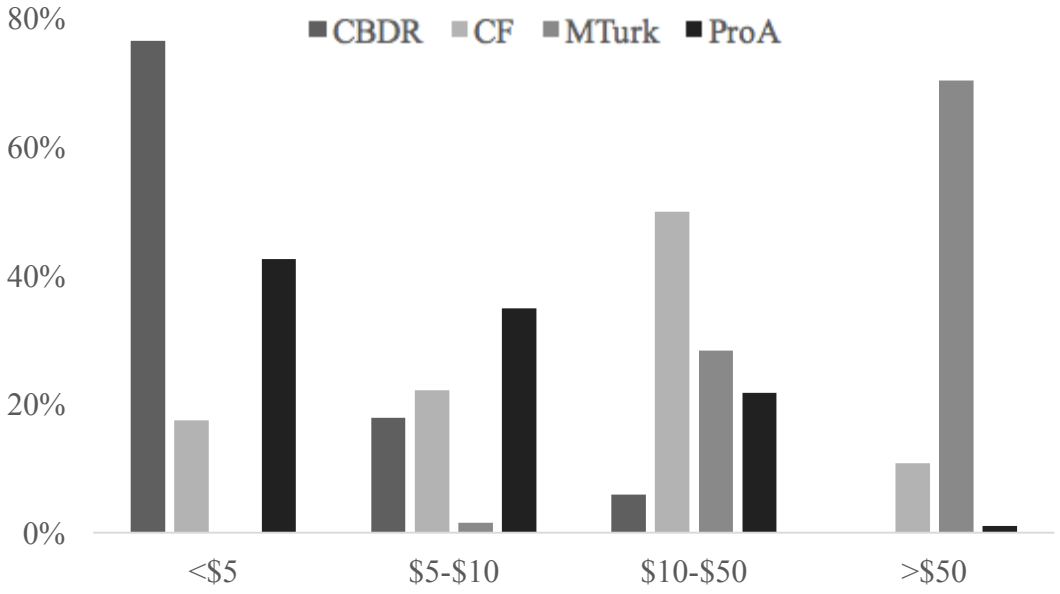


Figure 9. Percentage of participants in different quartiles of average weekly earning between the platforms (the cutoffs represent the quartiles of earnings in the overall sample).



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.

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

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 & Paollacci, 2014; Peer et al., 2013), 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 (less than 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 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 toward 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.

Study 2

Method

Samples' composition and characteristics. We recruited 1,374 participants from both sites (691 from MTurk and 683 from ProA), of which 1,205 (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. Particiapnts 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 MTuk 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, $p < .01$). We found statistically significant

differences between the sites in ethnicity, χ^2 (5) = 25.51, $p < .01$, education, χ^2 (9) = 60.04, $p < .01$, and income, χ^2 (7) = 147.02, $p < .01$, but not in English proficiency, t (0.05), $p = .96$ (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, $p < .01$. 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).

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 seconds. 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

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

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 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.

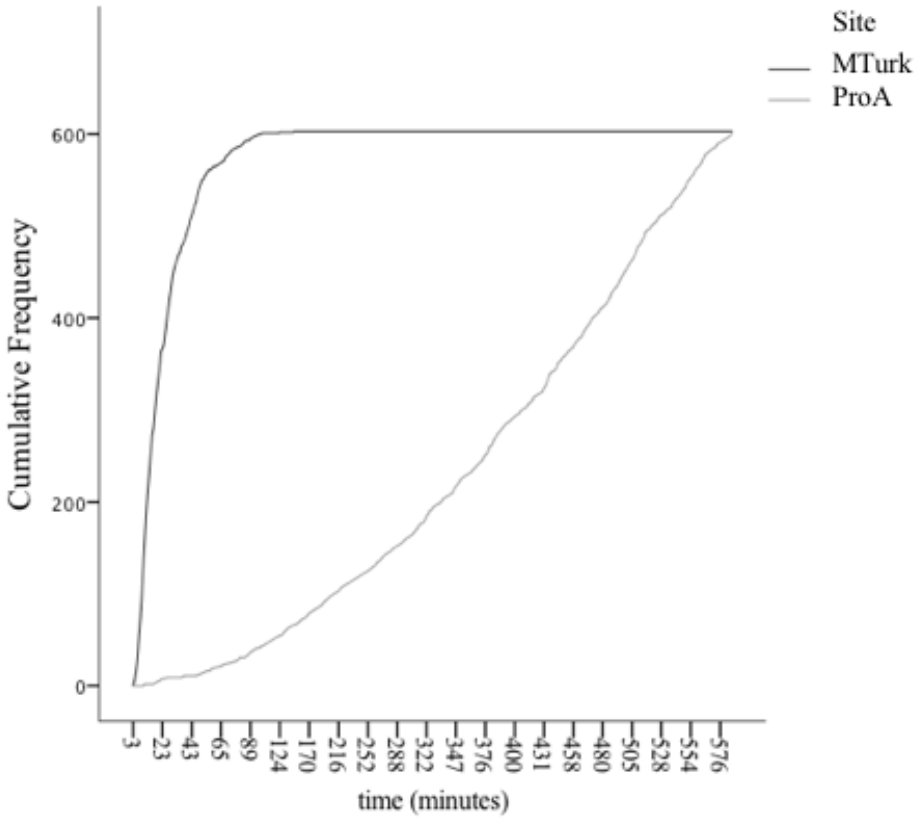
Participants then completed demographic and usage-related questions similar to Study 1. We also included questions that were designed to test some other hypotheses (for example, we asked participants how many shoes they owned in order to test the hypothesis that women have more shoes than men). The purpose 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.

Results

Response rates. As Figure 10 shows, the response rate on MTurk was much faster than on ProA in this study. While data collection was completed on MTurk in 151 minutes, it took almost 10 hours 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).

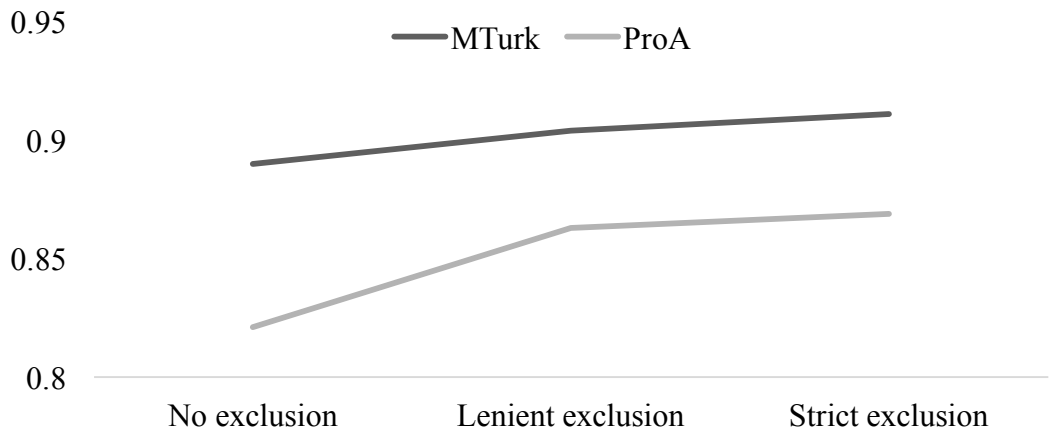
Figure 10. Response rates between the two platforms.



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, $p < .01$, 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, $p <$ .01

Figure 11. 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).

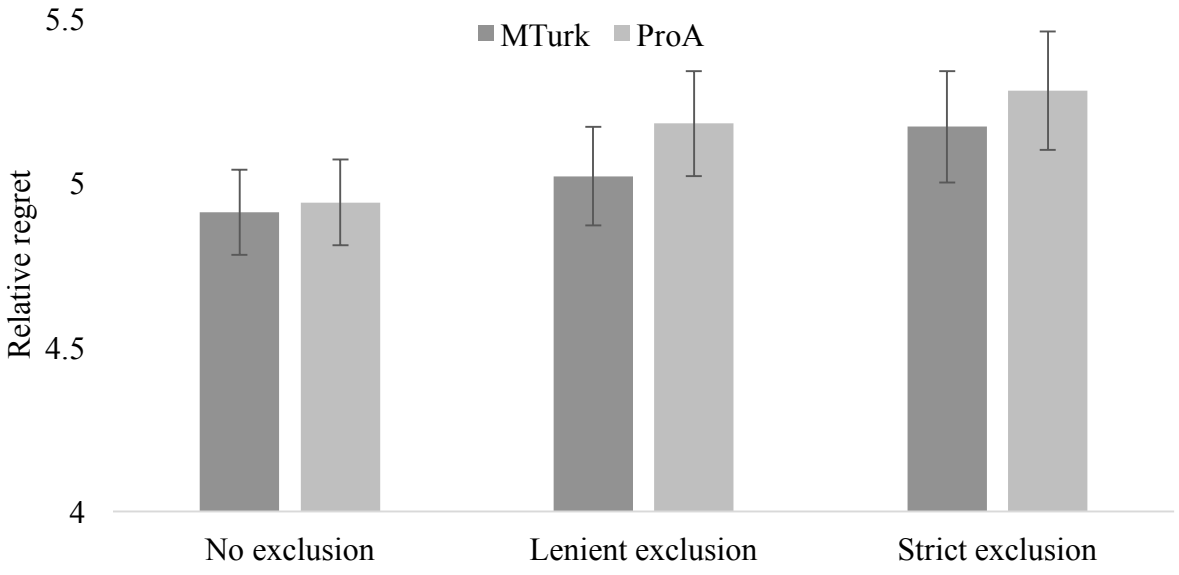


Reliability. After coding the CFC questions according to the scale, we examined its reliability between the sites and between exclusion policies. As can be seen in Figure 11, reliability on MTurk was found to be slightly higher than on ProA when using all participants (Cronbach's alpha = 0.89 vs. 0.821). 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 Hakistan and

Whalen's (1976) method, we found that the differences between all reliability coefficients were statistically significant (χ^2 (7) = 153.58, $p < .001$). However, it should be noted that in all instances reliability remained above the conventional threshold of 0.8 indicating adequate reliability.

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 Figure 12 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 differences were not statistically significant, as is evident by the overlap of the 95% confidence interval bars in Figure 12.

Figure 12. 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).



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.

Usage patterns. As can be seen in Figure 13, most MTurk participants reported spending between 8 and 20 hours per week on the platform. ProA users spent considerably less time, most reporting spending between 1 and 2 hours per week only. As Figure 14 shows, this clearly results in earning differences between the platforms: whereas more than 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 (5,900), 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.

Figure 13. Distribution of frequency of usage between the platforms.

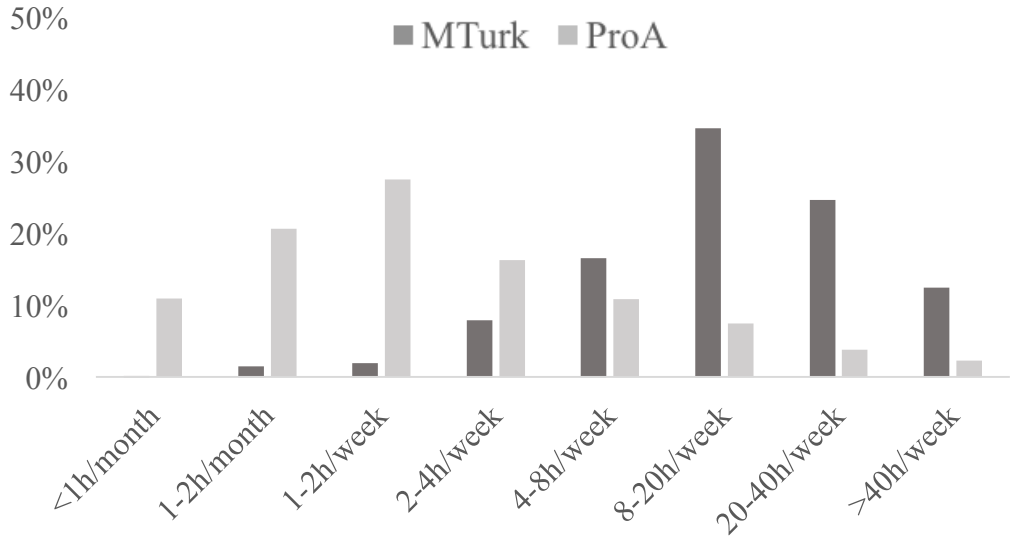
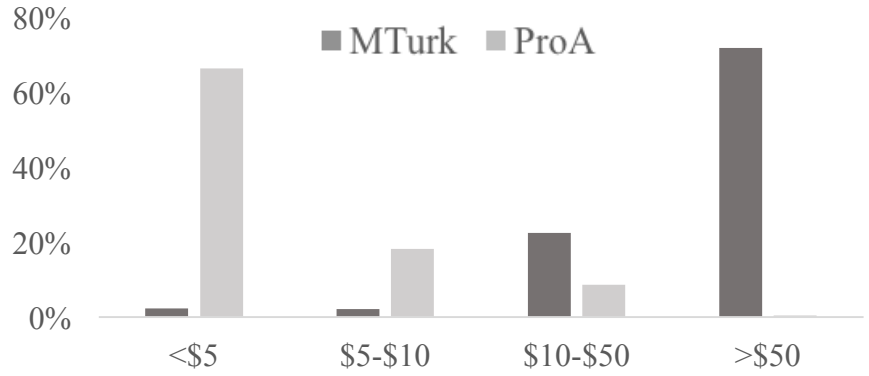


Figure 14. Distribution of average weekly earning between the platforms.



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 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% + 10p 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.

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., & Feng Kao, C. (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), 29-29.

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 Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology*.

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

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.

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. (2013). 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), 742-75

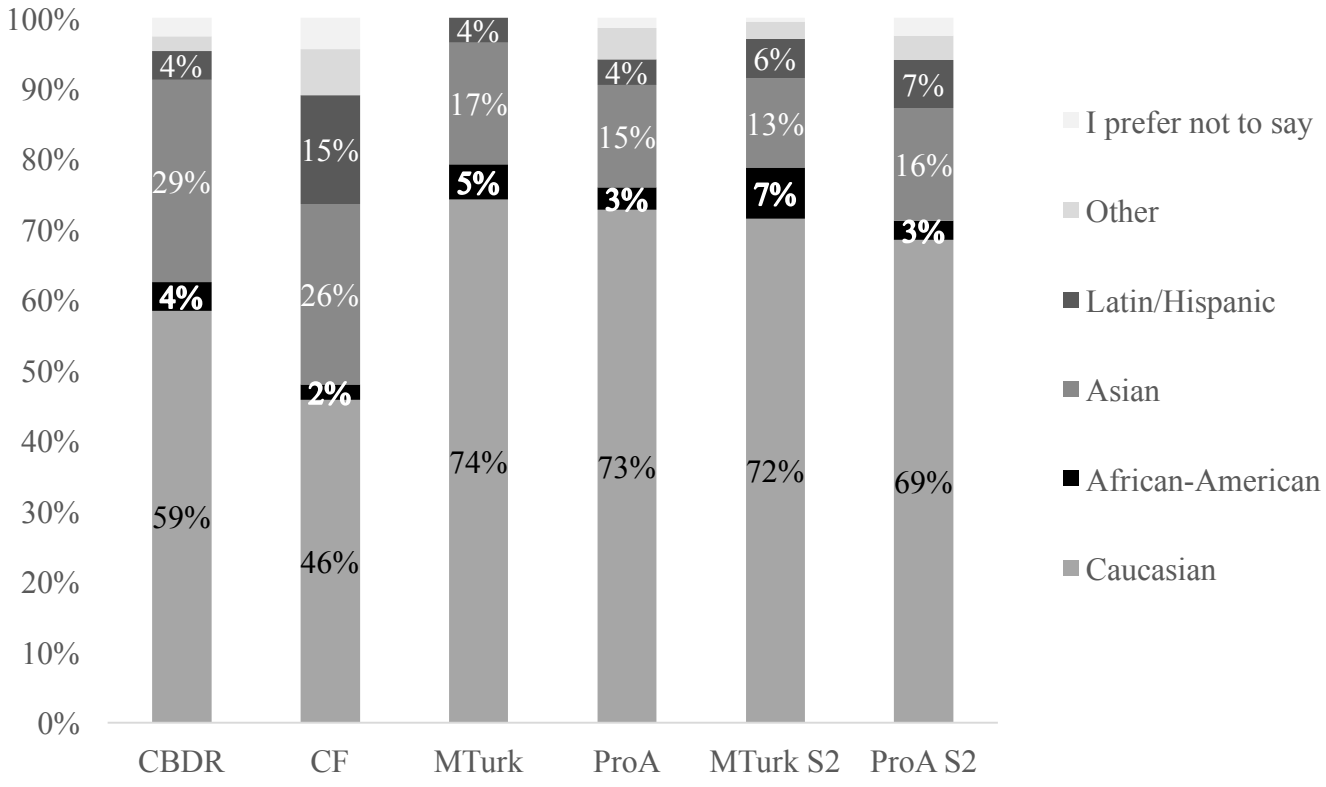
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

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.

Appendix – Additional figures.

Figure A1. Ethnicity distributions, from Studies 1 and 2 (S2 refers to Study 2)



Note: the same categories and labels were used on all platforms.

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

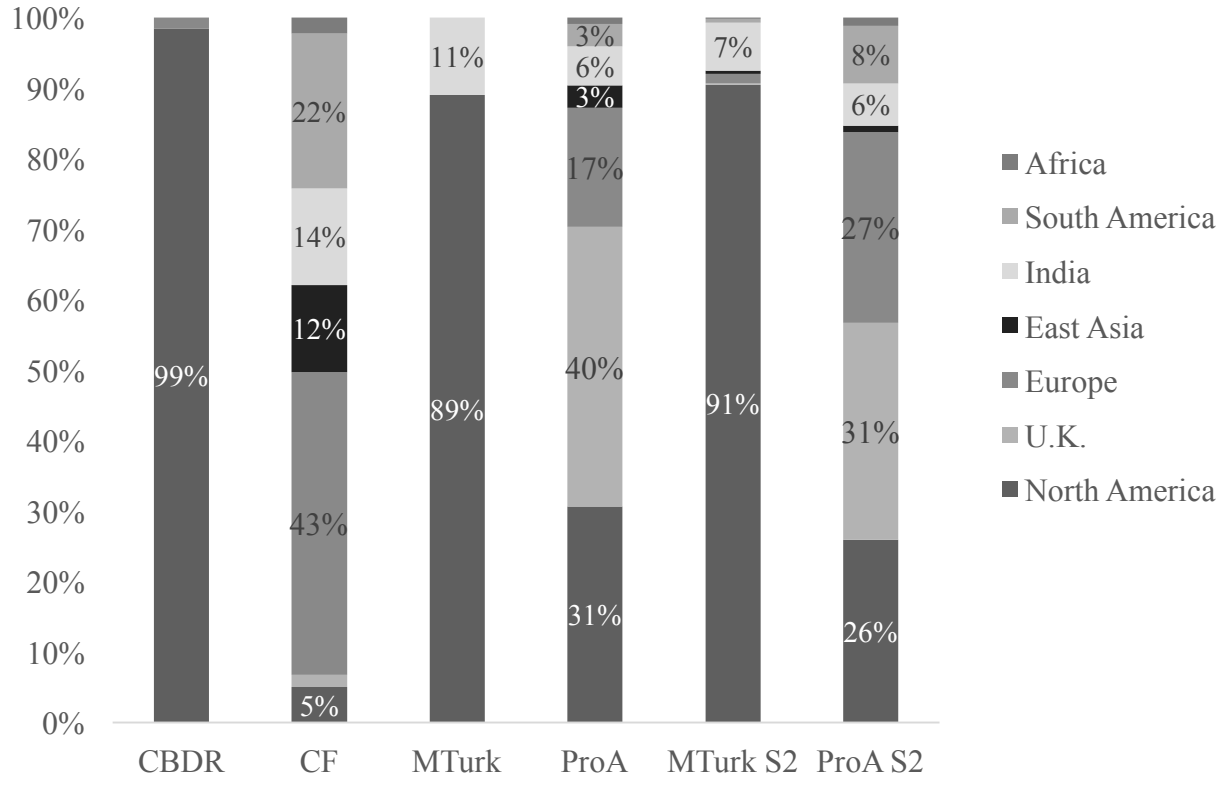


Figure A3. Reported income distributions from Studies 1 and 2 (S2 refers to Study 2).

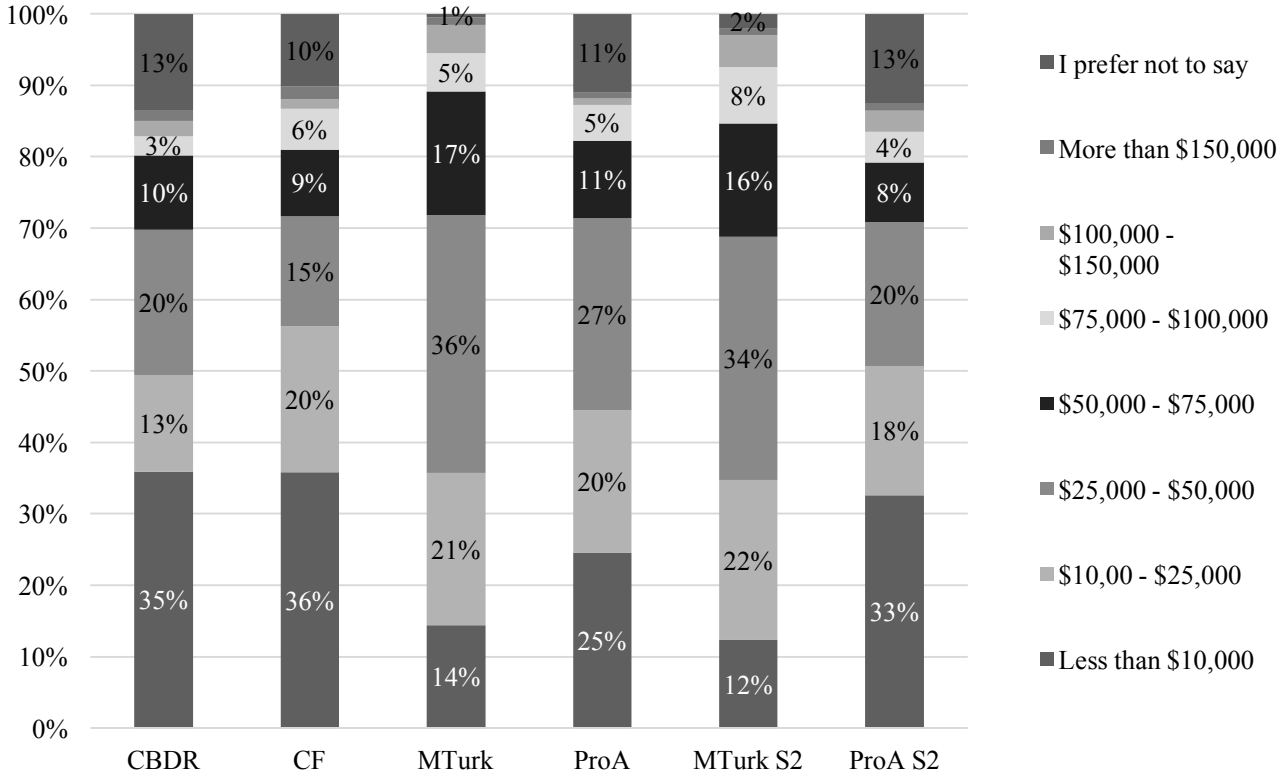
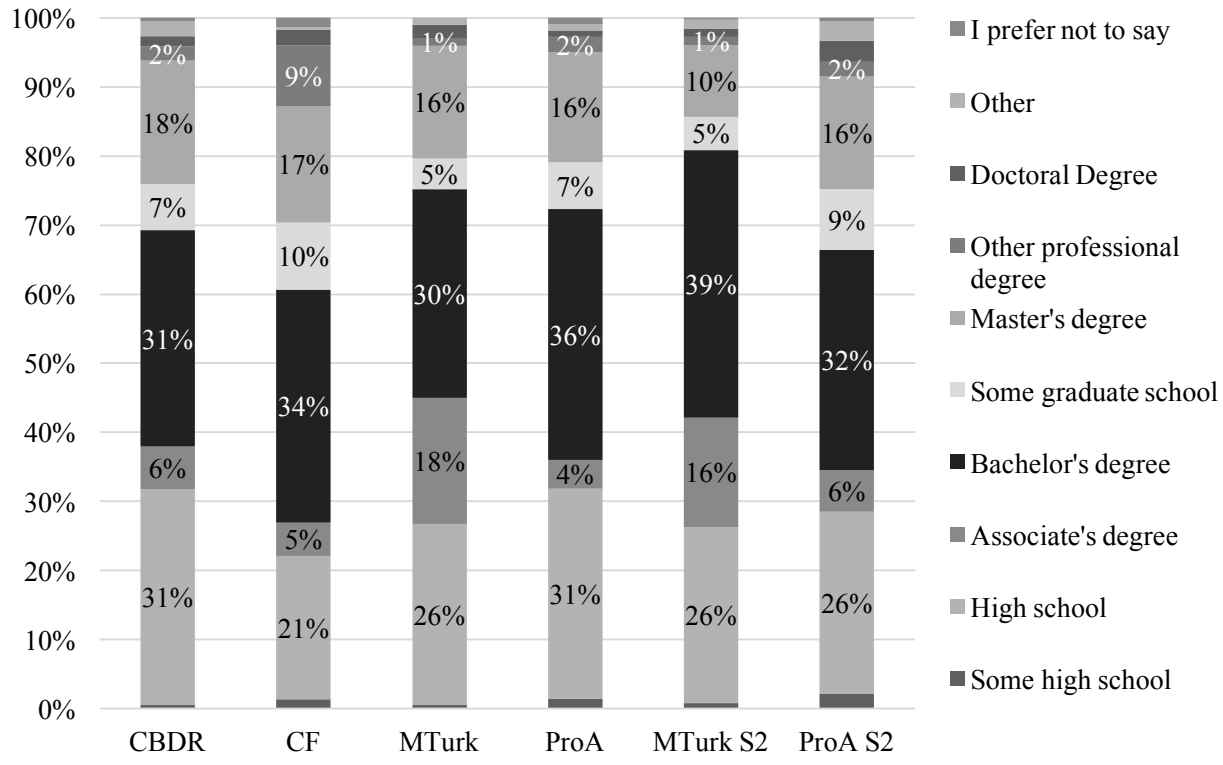


Figure A4. Reported education level distributions from Studies 1 and 2 (S2 refers to Study 2).



BeyondTurkStudy1

Q2 Hello and welcome to our Research Project! This survey is part of a research project we are conducting.The purpose is to better understand individual differences in personal attitudes, opinions and behaviors. The entire survey should take around 15 minutes of your time. To be eligible, you must be 18 years old or older. Participation in this research is completely voluntary, anonymous and confidential. If you agree to participate, please click "next."

Q76 This study is made out of many quick mini studies that you will complete. Many of the tasks are completely different from each other, so don't worry if some of the questions seem out of place. For many of the tasks, you will simply answer a question or complete a questionnaire, while for others you may be asked to write down your thoughts, and for others you may be asked to rapidly categorize words or pictures. Please pay careful attention to the instructions and answer as candidly as you can. Your participation is important to the research process and we greatly appreciate it.

Q4 This study requires you to voice your opinion using the scales below. It is important that you take the time to read all instructions and that you read questions carefully before you answer them. Previous research on preferences has found that some people do not take the time to read everything that is displayed in the questionnaire. The questions below serve to test whether you actually take the time to do so. Therefore, if you read this, please answer 'two' on the first question, add three to that number and use the result as the answer on the second question. Thank you for participating and taking the time to read all instructions.

Q6 I would prefer to live in a large city rather than a small city.

- 1-Strongly Disagree (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7-Strongly Agree (7)

Q8 I would prefer to live in a city with many cultural opportunities, even if the cost of living was higher.

- 1-Strongly Disagree (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7-Strongly Agree (7)

Q69 Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Q71 For research purposes, we are interested whether you are seeing parts of this experiment for the first time or not. So, after some of the questions we will ask you to indicate whether this was the first time you've been asked that question or not. Please answer honestly - you will get paid for the survey regardless of whether you've seen some questions before or not. So, regarding the previous question, was this the first time you've been asked to answer such a question?

- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q72 The next question measures your perceptual abilities. We will show you an image with several people in it. Some of the persons in the image will be clearly visible but some might be somewhat obscure. Your goal is to count the number of different persons you see in that image and to report it as quickly as possible. You will only have 20 seconds to observe the image and report your answer so please pay attention and answer carefully. As we've explained before, this survey is about individual differences and how different people react to different situations. Every person can be different, so we expect to get different results from different people. Please feel free to provide us with any response you personally think is appropriate, in the other parts of the survey. In this part, though, we ask that you ignore the instructions given above and when you see the image with the persons in it you must report you see zero persons in the picture, even if that is not correct. Thank you for following our instructions. Please click on next to proceed.

Q73

Q74 How many different persons can you see in the picture? Answer as quickly as possible.

- ❍ 0 (1)
- ❍ 1 (2)
- ❍ 2 (3)
- ❍ 3 (4)
- ❍ 4 (5)
- ❍ 5 (6)
- ❍ 6 (7)
- ❍ 7 (8)
- ❍ 8 (9)
- ❍ 9 (10)
- ❍ 10 (11)

Q75 Timing

First Click (1)
Last Click (2)
#QuestionText, TimingPageSubmit# (3)
#QuestionText, TimingClickCount# (4)

Q70 Regarding the previous question, was this the first time you've been asked to answer such a question?

- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q30 Following are several statements some people use to describe themselves. Please indicate how much you personally agree or disagree with each of the statements.

	Strongly Disagree (1)	Disagree (3)	Neither (3)	Agree (4)	Strongly Agree (5)
I feel that I am a person of worth, at least on an equal plane with others. (RSES_1)	❍	❍	❍	❍	❍
I feel that I have a number of good qualities.. (RSES_2)	❍	❍	❍	❍	❍
All in all, I am inclined to feel that I am a failure. (RSES_3)	❍	❍	❍	❍	❍
I am able to do things as well as most other people. (RSES_4)	❍	❍	❍	❍	❍
I feel I do not have much to be proud of. (RSES_5)	❍	❍	❍	❍	❍
I take a positive attitude toward myself. (RSES_6)	❍	❍	❍	❍	❍
I have never used the Internet myself (Q30_14)	❍	❍	❍	❍	❍
On the whole, I am satisfied with myself. (RSES_7)	❍	❍	❍	❍	❍
I wish I could	❍	❍	❍	❍	❍

have more respect for myself. (RSES_8)					
I certainly feel useless at times. (RSES_9)	❍	❍	❍	❍	❍
At times I think I am no good at all. (RSES_10)	❍	❍	❍	❍	❍

Q73 Regarding the previous questions, was this the first time you've been asked to answer such a questionnaire?

- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q45 Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows.Program A: If Program A is adopted, 200 people will be saved.Program B: If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

Q48 Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows.Program A: If Program A is adopted 400 people will die.Program B: If Program B is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

Q46 Which of the two programs would you favor?
- ❍ Program A (1)
- ❍ Program B (2)

Q74 Regarding the previous question, was this the first time you've been asked to answer such a question?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q33 Imagine that your favorite football team is playing an important game. You have a ticket to the game that you have paid handsomely for. However, on the day of the game, it happens to be freezing cold.

Q35 Imagine that your favorite football team is playing an important game. You have a ticket to the game that you have received for free from a friend. However, on the day of the game, it happens to be freezing cold.

Q34 What do you do?
- ❍ 1 - Definitely stay at home (1)
- ❍ 2 (2)
- ❍ 3 (3)
- ❍ 4 (4)
- ❍ 5 (5)
- ❍ 6 (6)
- ❍ 7 (7)
- ❍ 8 (8)
- ❍ 9 - Definitely go to the game (9)

Q75 Regarding the previous question, was this the first time you've been asked to answer such a question?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q78 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 6's. Based on your imagined scenario, how many times do you think the man had rolled the dice before you walked by?

Q79 Imagine that you are in a casino and you happen to pass a man rolling dice. You observe him roll three dice and one comes up 3, and two come up 6's. Based on your imagined scenario, how many times do you think the man had rolled the dice before you walked by?

Q80 How many times?

Q76 Regarding the previous question, was this the first time you've been asked to answer such a question?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q81 A quote similar to what follows was once spoken by George Washington: "I have sworn to only live free, even if I find bitter the taste of death."

Q83 A quote similar to what follows was once spoken by Osama Bin Laden: "I have sworn to only live free, even if I find bitter the taste of death."

Q84 How much do you agree with this quote?
- ❍ Strongly Disagree (15)
- ❍ Disagree (16)
- ❍ Somewhat Disagree (17)
- ❍ Neither Agree nor Disagree (18)
- ❍ Somewhat Agree (19)
- ❍ Agree (20)
- ❍ Strongly Agree (21)

Q77 Regarding the previous question, was this the first time you've been asked to answer such a question?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q14 For each of the statements below, please indicate to what extent the statement is characteristic of you by indicating how much you agree or disagree with each of them.

	Strongly Disagree (1)	Disagree (2)	Neither (3)	Agree (4)	Strongly Agree (5)
I would prefer complex to simple problems. (1)	❍	❍	❍	❍	❍
I like to have the responsibility of handling a situation that requires a lot of thinking. (2)	❍	❍	❍	❍	❍
Thinking is not my idea of fun (3)	❍	❍	❍	❍	❍
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities (4)	❍	❍	❍	❍	❍
I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something. (5)	❍	❍	❍	❍	❍
I find satisfaction in deliberating hard and for long hours. (6)	❍	❍	❍	❍	❍
I only think as	❍	❍	❍	❍	❍

hard as I have to (7)					
I prefer to think about small, daily projects to long-term ones (8)	❍	❍	❍	❍	❍
I like tasks that require little thought once I've learned them (9)	❍	❍	❍	❍	❍
The idea of relying on thought to make my way to the top appeals to me. (10)	❍	❍	❍	❍	❍
I really enjoy a task that involves coming up with new solutions to problems. (11)	❍	❍	❍	❍	❍
I currently don't pay attention to the questions I'm being asked in the survey (19)	❍	❍	❍	❍	❍
Learning new ways to think doesn't excite me very much (12)	❍	❍	❍	❍	❍
I prefer my life to be filled with puzzles that I must solve. (13)	❍	❍	❍	❍	❍
The notion of thinking	❍	❍	❍	❍	❍

abstractly is appealing to me. (14)					
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought. (15)	❍	❍	❍	❍	❍
I feel relief rather than satisfaction after completing a task that required a lot of mental effort (16)	❍	❍	❍	❍	❍
It's enough for me that something gets the job done; I don't care how or why it works (17)	❍	❍	❍	❍	❍
I usually end up deliberating about issues even when they do not affect me personally. (18)	❍	❍	❍	❍	❍

Q72 Regarding the previous questions, was this the first time you've been asked to answer such a questionnaire?

- ❍ Yes (1)
- ❍ No (2)
- ❍ Not sure (3)

Q18 Demographics it would be helpful to our research to better understand our participants' demographics. Please answer these questions as truthfully as possible.

Q20 What is your gender?

- ❍ Male (1)
- ❍ Female (2)

Q22 How old are you?

Q69 What is your ethnicity?

- ❍ Caucasian (1)
- ❍ African-American (2)
- ❍ Asian (3)
- ❍ Latin/Hispanic (4)
- ❍ Other (5) _____
- ❍ I prefer not to say (6)

Q24 What is the highest level of education that you have completed?

- ❍ Some high school (1)
- ❍ High school (2)
- ❍ Associate's degree (3)
- ❍ Bachelor's degree (4)
- ❍ Some graduate school (5)
- ❍ Master's degree (6)
- ❍ Other professional degree (7)
- ❍ Doctoral Degree (8)
- ❍ Other (9) _____
- ❍ I prefer not to say (10)

Q26 What is your annual income?

- ❍ Less than $10,000 (1)
- ❍ $10,00 - $25,000 (2)
- ❍ $25,000 - $50,000 (3)
- ❍ $50,000 - $75,000 (4)
- ❍ $75,000 - $100,000 (5)
- ❍ $100,000 - $150,000 (6)
- ❍ More than $150,000 (7)
- ❍ I prefer not to say (8)

Q32 In which country do you currently live?

Q30 What is your nationality?

Q73 How well can you read and write in each of the following languages?

	Not at all (1)	A little (2)	Good (3)	Very good (4)	Excellent (5)
English (1)	❍	❍	❍	❍	❍
French (2)	❍	❍	❍	❍	❍
German (3)	❍	❍	❍	❍	❍
Hindu/Tamil (4)	❍	❍	❍	❍	❍
Chinese/Mandarin (5)	❍	❍	❍	❍	❍
Japanese (6)	❍	❍	❍	❍	❍
Arabic (7)	❍	❍	❍	❍	❍

Q27 For how long have you been an active member of ${e://Field/site}?

- ❍ A few days (1)
- ❍ Less than a month (2)
- ❍ A few months (3)
- ❍ Less than a year (4)
- ❍ About a year (5)
- ❍ About two years (6)
- ❍ More than two years (7)

Q28 How often do you use ${e://Field/site}?

- ❍ Less than 1 hour per month (1)
- ❍ 1-2 hours per month (2)
- ❍ 1-2 hours per week (3)
- ❍ 2-4 hours per week (4)
- ❍ 4-8 hours per week (5)
- ❍ 8-20 hours per week (6)
- ❍ 20-40 hours per wekk (7)
- ❍ More than 40 hours per week (8)

Q29 Why do you use ${e://Field/site}?(check all that apply)

- ❑ It is my main source of income (1)
- ❑ To earn some extra money (2)
- ❑ To pass the time (3)
- ❑ Other (4) _____

Q63 What types of tasks do you usually do on ${e://Field/site}?

Q59 How much, on average, do you earn in a week using ${e://Field/site}?

Q64 How many tasks approximately, have you completed on ${e://Field/site}?

Q71 What percent of your submissions have been approved so far on ${e://Field/site}?

Q57 Have you used any of the web sites listed below and, if so, how often?

	Never (1)	Once or twice (2)	A few times (3)	Several times (4)	Many times (5)
ClickWorker (3)	❍	❍	❍	❍	❍
CrowdFlower (5)	❍	❍	❍	❍	❍
Amazon MTurk (8)	❍	❍	❍	❍	❍
Prolific Academic (9)	❍	❍	❍	❍	❍

Q58 Are you an active member in any other kind of crowd-sourcing web site or an online participants pool?

Q42 Thank you for your responses so far. As we promised, you will receive an additional bonus for completing the entire study. The following instructions explain how your bonus will be determined. Please read them carefully.

Q43 To determine your bonus, we will roll a six-sided die using our survey software. Your bonus will be the outcome of the die multiplied by 10 cents. For example, for the outcome 1, you will get 10 cents, 2 = 20 cents, 3 = 30 cents, etc.As you know, a die has 6 sides (with numbers from 1 to 6), each with an equal probability (of about 17%). Before we roll the die, we will ask you to choose whether you'd like to base your bonus on the number shown on the top ("up side") or the bottom ("down side"). As you know, the numbers on opposite sides of a die always sum up to 7. So, if you see 1 on the up side, 6 will be on the down side; if 4 is on the up side, 3 is on the down side, etc. We will ask you to choose a side (up or down) in your mind, before we roll the die. Then, we will roll the die and your bonus will be determined based on either the number shown (if you chose the up side) or the number on the opposite side (if you chose the down side). For example, if the die shows a "3", you will get 30 cents if you chose the up side or 40 cents if you chose the down side. Are these instructions clear? Please summarize them briefly in your own words.

Q70 Timing

- First Click (1)
- Last Click (2)
- Page Submit (3)
- Click Count (4)

Q44 OK. Now we will roll the die. Before you click on Next, please choose (in your mind) whether you'd like to use the up or down side of the die. When you are ready, click on next.

Q45 rolling die..please wait...

Q54 Timing
- First Click (1)
- Last Click (2)
- Page Submit (3)
- Click Count (4)

Q46

Q47

Q48

Q49

Q50

Q51

Display This Question:
If 1 Is Displayed

Q71 Did you choose "up" or "down"?
- ❍ up - 10 cents bonus (1)
- ❍ down - 60 cents bonus (2)

Display This Question:
If 2 Is Displayed

Q72 Did you choose "up" or "down"?
- ❍ up - 20 cents bonus (1)
- ❍ down - 50 cents bonus (2)

Display This Question:
If 3 Is Displayed

Q73 Did you choose "up" or "down"?
- ❍ up - 30 cents bonus (1)
- ❍ down - 40 cents bonus (2)

Display This Question:
If 4 Is Displayed

Q74 Did you choose "up" or "down"?
- ❍ up - 40 cents bonus (1)
- ❍ down - 30 cents bonus (2)

Display This Question:
If 5 Is Displayed

Q75 Did you choose "up" or "down"?
- ❍ up - 50 cents bonus (1)
- ❍ down -20 cents bonus (2)

Display This Question:
If 6 Is Displayed

Q76 Did you choose "up" or "down"?
- ❍ up - 60 cents bonus (1)
- ❍ down - 10 cents bonus (2)

Q77 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Q79 If you have any last comment/questions, please enter them below.

Display This Question:
If site Is Equal to prolific
Or site Is Equal to crowdflower

Q73 To complete the study, please enter your Participant ID and click on next.

BeyondTurkStudy2

Q1.1 Hello and welcome to our Research Project! This survey is part of a research project we are conducting.The purpose is to better understand individual differences in personal attitudes, opinions and behaviors. The entire survey should take about 15-20 minutes of your time. To be eligible, you must be 18 years old or older. Participation in this research is completely voluntary, anonymous and confidential. If you agree to participate, please click "next."

Q1.2 This study contains multiple mini-tasks and questions that may, or may not, be related to each other. The tasks appear in random order so please try to disregard previous tasks when you're doing the current one. Click next to begin.

Q2.1 The next question measures your perceptual abilities. We will show you an image with several people in it. You will only have 10 seconds to count and report how many people you see. Click next to see the image. As we've explained before, this survey is about individual differences and how different people react to different situations. Every person can be different, so we expect to get different results from different people. Please feel free to provide us with any response you personally think is appropriate, in the other parts of the survey. In this part, though, we ask that you ignore the instructions given above and when you see the image with the persons in it you must report you see zero persons in the picture, even if that is not correct. In the other part of the survey, please keep reading instructions carefully and answer questions candidly and truthfully. Please click on next to proceed.

Q2.2

Q2.3 How many people can you see in the picture? Answer as quickly as possible.

Q2.4 Timing

- First Click (1)
- Last Click (2)
- Page Submit (3)
- Click Count (4)

Q3.1 In this task you're asked to list your liking of the activities described below. It is important that you take the time to read all instructions and that you read questions carefully before you answer them. Research on how people fill out surveys has found that some people do not take the time to read everything that is displayed in the questionnaire. The questions below serve to test whether you actually take the time to do so. Therefore, if you read this, please answer 'six' on the first question, divide that number by two and use the result as the answer on the second and third question. Thank you for participating and taking the time to read all instructions. Please answer the questions below.

	1-not at all (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9-Very much (9)
How much do you like math? (4)	❍	❍	❍	❍	❍	❍	❍	❍	❍
How much do you enjoy working with numbers? (5)	❍	❍	❍	❍	❍	❍	❍	❍	❍
How much do you enjoy thinking about difficult problems? (6)	❍	❍	❍	❍	❍	❍	❍	❍	❍

Q4.1 For each of the statements below, please indicate whether or not the statement is characteristic of you. If the statement is extremely uncharacteristic of you (not at all like you) please fill-in a "1"; if the statement is extremely characteristic of you (very much like you) please fill-in a "5". Use the numbers in the middle if you fall between the endpoints. Please keep the following scale in mind as you rate each of the statements below.

	1-extremely uncharacteristic (1)	2-somewhat uncharacteristic (2)	3-uncertain (3)	4-somewhat characteristic (4)	5-extremely characteristic (5)
I consider how things might be in the future, and try to influence those things with my day to day behavior. (1)	❍	❍	❍	❍	❍
Often I engage in a particular behavior in order to achieve outcomes that may not result for many years. (4)	❍	❍	❍	❍	❍
I only act to satisfy immediate concerns, figuring the future will take care of itself. (7)	❍	❍	❍	❍	❍
My behavior is only influenced by the immediate (i.e., a matter of days or weeks) outcomes of my actions. (10)	❍	❍	❍	❍	❍
My convenience is a big factor in the decisions I	❍	❍	❍	❍	❍

make or the actions I take. (11)					
I am willing to sacrifice my immediate happiness or well-being in order to achieve future outcomes. (12)	❍	❍	❍	❍	❍
I think it is important to take warnings about negative outcomes seriously even if the negative outcome will not occur for many years. (9)	❍	❍	❍	❍	❍
I think it is more important to perform a behavior with important distant consequences than a behavior with less important immediate consequences. (8)	❍	❍	❍	❍	❍
I think I have never used the Internet myself at any time through the course of my personal life. (16)	❍	❍	❍	❍	❍
I generally ignore warnings about possible future problems	❍	❍	❍	❍	❍

because I think the problems will be resolved before they reach crisis level. (6)					
I think that sacrificing now is usually unnecessary since future outcomes can be dealt with at a later time. (5)	❍	❍	❍	❍	❍
I only act to satisfy immediate concerns, figuring that I will take care of future problems that may occur at a later date. (3)	❍	❍	❍	❍	❍
Since my day to day work has specific outcomes, it is more important to me than behavior that has distant outcomes. (13)	❍	❍	❍	❍	❍

Q5.1 The next part of the study is a memory task. In the following, we ask that you try to recall four (4) events from your personal life in the past year that have made you feel happy. Please describe each of the four (4) happy events in the fields below.

happy event #1 (1)

happy event #2 (2)

happy event #3 (3)

happy event #4 (4)

Q5.2 The next part of the study is a memory task. In the following, we ask that you try to recall twelve (12) events from your personal life in the past year that have made you feel happy. Please describe each of the twelve (12) happy events in the fields below.

happy event #1 (1)
happy event #2 (2)
happy event #3 (3)
happy event #4 (4)
happy event #5 (5)
happy event #6 (6)
happy event #7 (7)
happy event #8 (8)
happy event #9 (9)
happy event #10 (10)
happy event #11 (11)
happy event #12 (12)

Q5.4 How happy or unhappy do you feel right now?

- ❍ Extremely happy (24)
- ❍ Moderately happy (25)
- ❍ Slightly happy (26)
- ❍ Neither happy nor unhappy (27)
- ❍ Slightly unhappy (28)
- ❍ Moderately unhappy (29)
- ❍ Extremely unhappy (30)

Q5.3 How difficult did you find this memory task to be?

- ❍ Extremely easy (10)
- ❍ Moderately easy (11)
- ❍ Slightly easy (12)
- ❍ Neither easy nor difficult (13)
- ❍ Slightly difficult (14)
- ❍ Moderately difficult (15)
- ❍ Extremely difficult (16)

Q55 Please read the following description carefully and answer the question below. 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 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. Who do you think felt her dawdling was more foolish (or irresponsible), Mrs. Crane or Mrs. Tees?

- ❍ Mrs. Crane felt more foolish considerably (1)
- ❍ Mrs. Crane felt more foolish moderately (2)
- ❍ Mrs. Crane felt more foolish slightly (3)
- ❍ They felt foolish similarly (4)
- ❍ Mrs. Tees felt more foolish slightly (5)
- ❍ Mrs. Tees felt more foolish moderately (6)
- ❍ Mrs. Tees felt more foolish considerably (7)

Q56 For the next part of the study we'd like to ask you some questions about yourself. Please read each question carefully and answer it as candidly as possible. Questions are presented in random order.

Q6.23 How close are you to retirement age?

- ❍ I'm retired (0)
- ❍ Very close (1)
- ❍ Close (2)
- ❍ Far (3)
- ❍ Very far (4)
- ❍ Very far away (5)
- ❍ I prefer not to say (99)

Q6.24 How many pairs of shoes do you own, approximately?

Q6.25 How important is social equality to you, personally?

- ❍ Extremely important (5)
- ❍ Very important (4)
- ❍ Moderately important (3)
- ❍ Slightly important (2)
- ❍ Not at all important (1)

Q6.26 How likely do you think are people to die of smoking?
- ❍ Extremely likely (5)
- ❍ Somewhat likely (4)
- ❍ Neither likely nor unlikely (3)
- ❍ Somewhat unlikely (2)
- ❍ Extremely unlikely (1)

Q6.1 Usage questions We'd like to know more about how you use online web sites to take part in research and surveys. Please answer the following questions on that topic.

Q6.2 For how long have you been an active member of ${e://Field/site}?
- ❍ A few days (1)
- ❍ Less than a month (2)
- ❍ A few months (3)
- ❍ Less than a year (4)
- ❍ About a year (5)
- ❍ About two years (6)
- ❍ More than two years (7)

Q6.3 How often do you use ${e://Field/site}?
- ❍ Less than 1 hour per month (1)
- ❍ 1-2 hours per month (2)
- ❍ 1-2 hours per week (3)
- ❍ 2-4 hours per week (4)
- ❍ 4-8 hours per week (5)
- ❍ 8-20 hours per week (6)
- ❍ 20-40 hours per wekk (7)
- ❍ More than 40 hours per week (8)

Q6.4 Why do you use ${e://Field/site}?(check all that apply)
- ❑ It is my main source of income (1)
- ❑ To earn some extra money (2)
- ❑ To pass the time (3)
- ❑ Other (4) _____

Q6.5 How much, on average, do you earn in a week using ${e://Field/site}?

Q6.6 How many tasks approximately, have you completed on ${e://Field/site}?

Q6.7 What percent of your submissions have been approved so far on ${e://Field/site}?

Q6.8 Have you used any of the web sites listed below and, if so, how often?

	Never (1)	Once or twice (2)	A few times (3)	Several times (4)	Many times (5)
CrowdFlower (5)	❍	❍	❍	❍	❍
Amazon MTurk (8)	❍	❍	❍	❍	❍
Prolific Academic (9)	❍	❍	❍	❍	❍
Other (10)	❍	❍	❍	❍	❍

Display This Question:

If Have you used any of the web sites listed below and, if so, how often? Other Is Greater Than 1

Q6.9 What's the name of the other site you use?

Q6.10 Demographics it would be helpful to our research to better understand our participants' demographics. Please answer these questions as truthfully as possible.

Q6.11 What is your gender?
- ❍ Male (1)
- ❍ Female (2)

Q6.12 How old are you?

Q6.13 What is your height?

Q6.14 What is your weight?

Q6.15 Do you smoke cigarettes?
- ❍ Yes (1)
- ❍ No (3)
- ❍ I prefer not to say (4)

Q6.17 What is your ethnicity?
- ❍ Caucasian (1)
- ❍ African-American (2)
- ❍ Asian (3)
- ❍ Latin/Hispanic (4)
- ❍ Other (5) _____
- ❍ I prefer not to say (6)

Q6.18 What is the highest level of education that you have completed?
- ❍ Some high school (1)
- ❍ High school (2)
- ❍ Associate's degree (3)
- ❍ Bachelor's degree (4)
- ❍ Some graduate school (5)
- ❍ Master's degree (6)
- ❍ Other professional degree (7)
- ❍ Doctoral Degree (8)
- ❍ Other (9) _____
- ❍ I prefer not to say (10)

Q6.19 What is your annual income?
- ❍ Less than $10,000 (1)
- ❍ $10,00 - $25,000 (2)
- ❍ $25,000 - $50,000 (3)
- ❍ $50,000 - $75,000 (4)
- ❍ $75,000 - $100,000 (5)
- ❍ $100,000 - $150,000 (6)
- ❍ More than $150,000 (7)
- ❍ I prefer not to say (8)

Q6.20 In which country do you currently live?

Q6.21 What is your nationality?

Q6.16 How would you describe your political affiliation?
- ❍ Liberal or Left-wing (1)
- ❍ Conservative or right-wing (2)
- ❍ other (3) _____
- ❍ I prefer not to say (4)

Q6.22 How well can you read and write in English?

- ❍ Extremely well (5)
- ❍ Very well (4)
- ❍ Moderately well (3)
- ❍ Slightly well (2)
- ❍ Not well at all (1)
- ❍ I prefer not to say (99)

Q7.1 If you have any last comment/questions, please enter them below.

Display This Question:
If site Is Equal to prolific
Or site Is Equal to crowdflower

Q7.2 To complete the study, please enter your Participant ID and click on next.

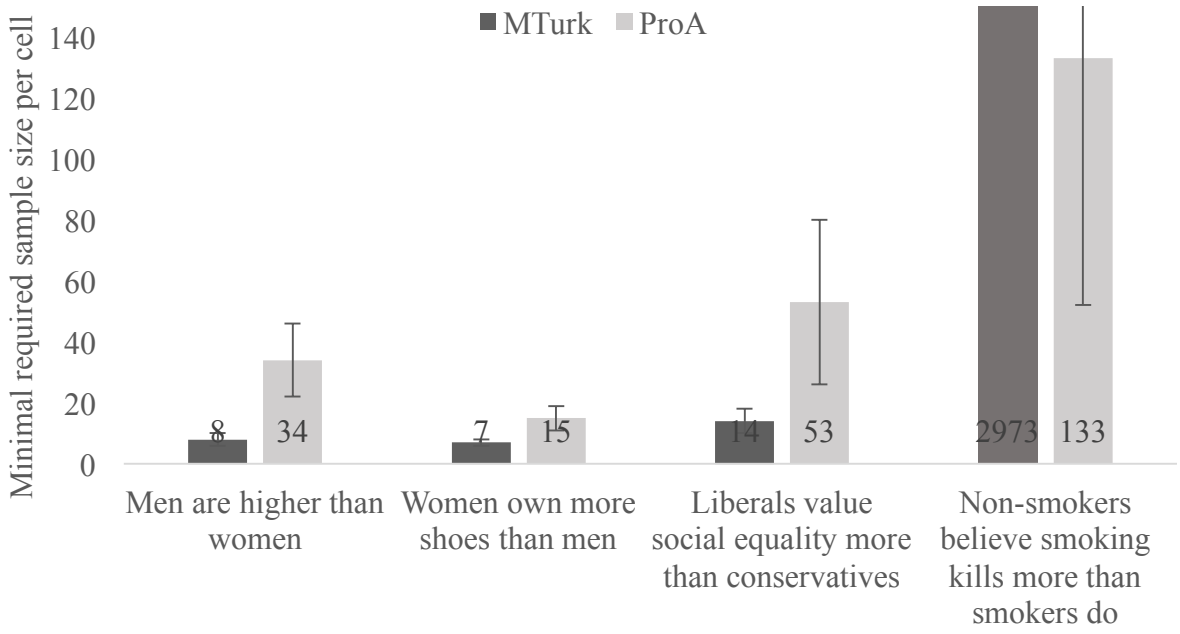
Beyond the Turk: Alternative platforms for crowdsourcing behavioral research

Supplementary file - Power tests

In another attempt to examine reproducibility, we adopted the approach of Simmons, Nelson, & Simonsohn (2013), and included in the study several questions that were aimed to test four relatively mundane hypotheses: a) that men are taller than women; b) that women own more shoes than men; c) that liberals think social equality is more important than conservatives; d) that smokers think that smoking kills less than non-smokers do. Following Simmons (2013), our goal with including these questions was to be able to determine what is the smallest sample size required (with a power of 80%) to arrive at a statistically significant result for each of these hypotheses. By comparing these minimal required sample sizes for the hypotheses between the samples, we aimed to measure another aspect of the quality of the data obtained from each sample.

Following Simmons et al. (2013), we examined the differences in the minimal sample sizes required to show a hypothesis is statistically significant, given 80% statistical power. To do that, we followed their recommended steps (Simmons, 2016, personal correspondence): First, the effect sizes for each hypothesis was computed on each sample (MTurk and ProA). Then, following Wuensch (2012), we computed the 95% confidence interval for each effect size. Finally, using a power calculator, we computed the minimal required sample size (per cell) needed to reach a $p < .05$ result with 80% statistical power for each of the effect sizes and their confidence intervals. The results are given in Figure S1. As can be seen, in almost all cases, the required sample on MTurk was smaller than ProA, except for the last hypothesis (that non-smokers believe smoking kills more than smokers do) because that hypothesis was not statistically significant among the MTurk sample, t (592) = 0.69, p = 0.49.

Figure S1. Minimal required sample size across platforms and across four hypotheses.



References

Simmons, J., Nelson, L. D., & Simonsohn, U. (2013). Life After P-Hacking. *Advances in Consumer Research*, 41

Wuensch, K. L. (2012). *Using SPSS to obtain a confidence interval for Cohen's d*. Retrieved October 12, 2016, from http://core.ecu.edu/psyc/wuenschk/SPSS/CI-d-SPSS.pdf .

Preliminary study report:

Alternative platforms for crowdsourcing online behavioral research

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

[a] Corresponding author. Graduate School of Business Administration, Bar-Ilan University, Ramat-Gan, 52900, Israel. eyal.peer@biu.ac.il

[b] Heinz College, Carnegie Mellon University, Pittsburgh, PA.

This report contains the details of a preliminary study conducted during a project that has been summarized in the paper "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research"

The purpose of this preliminary study was to survey the data quality of several alternative platforms for crowdsourcing behavioral research online. By searching for crowdsourcing websites, we identified six potential platforms that seem similar to Mechanical Turk in general design and purpose: CrowdFlower, MicroWorkers, RapidWorkers, MiniJobz, ClickWorker and ShortTask. Table 1 compares the features provided by each of the six platforms. Unlike MTurk, some of these platforms require that a "task" posted by a "requester" (for instance, a survey posted by a researcher) be reviewed before it is made available to "workers" (that is, the study participants). Based on our experience, this review process can take up to 24 hours. However, ClickWorker was found to be different, as researchers are only given the option to submit their survey questions offline to the site's team, and that team then designs and administers the survey to its pool of participants (for pay, of course). Most of the sites we reviewed have supervision and reputation mechanisms designed to measure the quality of submitted work, and to allow researchers to invite workers with high reputation (similar to MTurk's "approval rating" system, e.g., Peer et al., 2014). These sites also offer researchers the ability to sample participants based on other characteristics, such as country and language. Across these sites, common tasks completed by workers include assessing data quality and classifying and categorizing information, similar to the tasks commonly found on MTurk. All of the sites also include surveys, and many of them include tasks that require users to sign-up for various other sites, to click on links, and download applications. All sites offer the ability to request participants to submit some sort of a completion code that is given to them at the end of the survey, so researchers can approve tasks only from participants who actually completed the task/survey. At the time of our data collection, only MicroWorkers provided the ability to pay bonus payments to

participants after they submit their work (since then, CrowdFlower has also added this functionality).

Method

Sampling and participants. We attempted to set up a survey on all six of the aforementioned platforms, but were successful in doing so only on three of them. MiniJobz rejected our survey after reviewing it, without providing any reason for rejection. Despite our repeated attempts to contact their team, we did not receive a response from them about why the survey was rejected. Clickworker asked for a price of more than $800 to set up our survey for 200 participants. Because this price is considerably higher than the cost of running the same survey on the other platforms, we opted against using this site. On ShortTask, we received a credit card error despite using a working credit card, and our repeated efforts to contact them were left unrequited.

We sampled 200 participants from each of the remaining platforms: CrowdFlower, MicroWorkers, RapidWorkers and MTurk. We also sampled participants from the Center for Behavioral Decision Research (CBDR) participant pool, which is a participant pool (including students and non-students) managed out of Carnegie Mellon University. We limited the sampling time to one week, in order to set a common timeframe for the study. During that week, we were able to sample about 200 participants from all platforms except for RapidWorkers, and ended up with a total sample of 890 participants. Table 2 shows the sample size obtained from each of these platforms, 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 sites within the period of late September to early October, 2014; surveys were submitted on week days during the morning hours (between 9am and 12pm EST); we did not set any restrictions (such as

location or previous ratings) on the survey on any of the sites, mainly because we wanted to be able to assess differences between the sites on these aspects. Due to their process, MicroWorkers and RapidWorkers published our survey about 24 hours after we submitted it, and CBDR published our survey in the evening of the same day, whereas our survey was started immediately on all other platforms.

Procedure. The survey included several stages. The first stage consisted of several psychometric questionnaires and experimental tasks adopted from prominent behavioral and psychological studies, which were administered in random order and designed to test attention, reliability, reproducibility, non-naivety, and individual differences. The second stage included a die-rolling task that tested dishonest behavior. The last stage included demographic and usage-related questions designed to better understand the different populations and their use of the different platforms. We next elaborate on the materials used in each stage.

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

Sample	Started the study	Completed	Percent of dropouts	Percent males	Mean age (SD)
MTurk	240	200	16.7%	59.0%	34.55 (11.13)
CrowdFlower	230	203	11.7%	72.1%	30.97 (10.04)
MicroWorkers	232	188	44.0%	50.5%	32.44 (11.76)
RapidWorkers	127	105	17.3%	87.4%	25.09 (4.44)
CBDR	215	194	9.8%	30.4%	30.78 (12.98)

Materials. To examine reliability of data and individual differences between platforms, we used three scales: the Need For Cognition scale (NFC, Cacioppo, Petty, & Kao, 1984), the Internet User Information Privacy Concerns scale (IUIPC, Malhotra, Kim, & Agarwal, 2004), 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 been previously used with success to measure data quality on MTurk (Peer et al., 2014). The IUIPC uses a scale from 1 (strongly disagree) to 7 (strongly agree) while the NFC and RSES use a scale from 1 (strongly disagree) to 5 (strongly agree). The order of the scales was randomized between participants.

To examine participants' attention, we used two attention-check questions (Peer et al., 2014). The first question came at the beginning and asked participants to respond to two questions on a 7-point scale: a) would you prefer living in a small or large city? b) Would you prefer to live in a city with many cultural opportunities, even if the cost of living was higher? In the instructions to these questions participants were asked not provide their actual responses and to select "two" for the first question, add three to that number and use the result to answer the second question. The second attention check question was embedded into the IUIPC scale as a statement that read, "I am not reading the questions in this survey."

To examine participants' non-naivety, which is their level of familiarity with commonly used research materials, we included the Cognitive Reflection Task, as Chandler et al. (2014) used to explore non-naivety among MTurk participants. The task included the original three-items version from Frederick (2005), as well as a newer three-items version used by Chandler et al. (2014). We hypothesized that naïve participants would solve the questions in this task less well than non-naïve participants, and thus a lower number of correct responses would (indirectly) suggest the existence of more naïve participants in that sample.

To examine reproducibility of known effects, we included three decision-making tasks. The first task was based on the Asian-disease framing effect (Tversky & Kahneman, 1981), in which participants were asked to imagine that the U.S. is preparing for the outbreak of a disease and select from two courses of action described either in positive (lives saved) frame or negative

(lives lost) frame: Program A, under which [200 people will be saved] [400 people will die] or Program B, under which there is a 1/3 probability that 600 people will be saved [no people will die] and 2/3 probability that no people will be saved [600 people will die]. The second task was based on the sunk-cost fallacy (following Oppenheimer et al., 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 an anchoring-and-adjustment task adopted from Jacowitz & Kahneman (1995), in which participants made four quantitative estimates (the length of the Mississippi River, the population of Chicago, the height of Mount Everest, and how many babies are born per day in the United States) after being told that the target is greater than (low-anchor condition) or less than (high-anchor condition) a specified value. The order of these tasks, as well as the questions within each task, were randomized between participants, and allocation to conditions was randomized within each of these tasks.

After participants completed all of the above parts, the next 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 in order to gain additional reward. Participants were told that the survey software would virtually roll a six-sided die and 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 up-facing number of the die, or the number opposite to it, facing down. This choice was to be made in their minds before the roll of the die. Then, the die was rolled (using a random number

generator that simulated a die roll) and participants were asked to report the number shown on the die and whether they picked the up or down 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, this task gave participants an incentive to cheat by declaring that they picked the down side when the side facing up showed a low number, and vice versa, that they picked the up side when the die roll showed a high number on that side. This task was employed only on the sites that allowed for post-completion bonuses: MTurk and MicroWorkers. For CBDR participants, the same task was used, but instead of 10 cents each number on the die gave the participant an extra raffle ticket to win a gift card worth $50.

At the end of the study, participants answered demographic questions and questions that pertained to the use of their respective site and other sites (which would be detailed in the following Results section).

Results

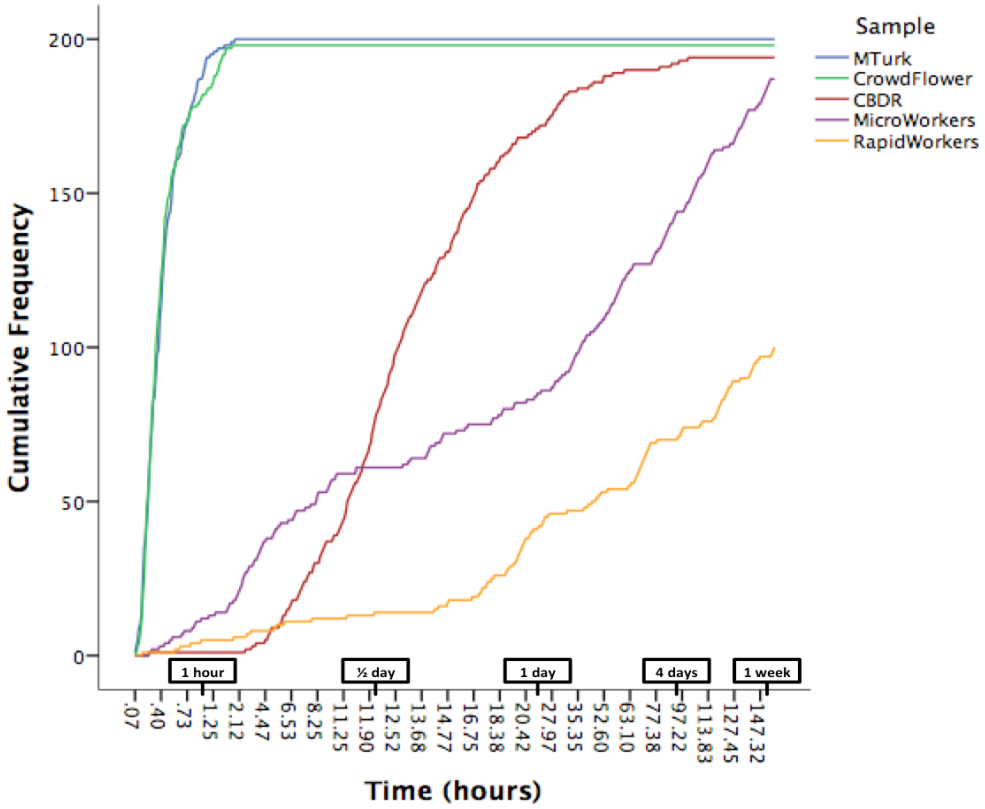
All our analyses include only participants who completed the entire study.

Response rates. As detailed in Table 1, dropout rates ranged between 10% (for CBDR) to about 17% (for RapidWorkers), except for MicroWorkers, where 44% of respondents that started the survey did not complete it. These dropout rates differed significantly between the five platforms, χ^2 (4) = 385.46, $p < .01$, and these differences remained significant even when excluding MicroWorkers, χ^2 (4) = 29.99, $p < .01$. Response rates also differed across sites. Figure 1 shows the cumulative frequency (absolute number) of accumulating responses according to the time (in hours) from the onset of the survey, counted from the start time of the first respondent for each sample until the end time of the last respondent for each sample (note that this does not include any delay between the time we posted our survey to the time the first respondent began taking it). As can be seen, MTurk and CrowdFlower showed the fastest

response rates, with all responses collected within less than 2 hours (101.01 and 108.55 responses per hour, respectively). With a considerable difference, CBDR showed the third fastest response rate (1.42 responses per hour).[1] MicroWorkers showed a slower response rate (1.08 responses per hour) and RapidWorkers, which failed to produce 200 responses, showed the slowest response rate (0.63 responses per hour). The hourly average response rates show that in order to collect 500 responses, one would have to wait only about 5 hours on either MTurk or CrowdFlower, about two weeks on CBDR, almost twenty days on MicroWorkers, and a full month on RapidWorkers.

Figure 1. Response rates across samples.

[1] This lower rate seems to have resulted from the fact that, although we submitted our survey in the morning, it was actually made available at around 11 pm on the same day. Apparently, only one participant completed it during that night and the others joined the following day.



To summarize, CrowdFlower provided a comparable alternative to MTurk in terms of response rates, and had a higher response rate than the CBDR university pool. MicroWorkers also had a satisfactory response rate, although somewhat slower than the university pool, and RapidWorkers's results were unsatisfactory. It should be noted that these differences in response rates may have originated from either the different size or quality of the population on the different sites, but may have also come from ancillary differences in technical aspects of our sampling, such as the exact timing of the submission and publication of the survey.

We also found differences in the time taken by participants from the different samples to complete our study. Because the time distribution was highly skewed, we compared medians across groups to find it was lowest on RapidWorkers (11.58 mins), followed by MTurk (13.06 mins), MicroWorkers and CBDR (16.77 and 16.88) and the highest on CrowdFlower (17.67 mins). A Kruskal-Wallis test showed these differences were statistically significant ($p < .01$).

Attention. Using two attention-check questions (Peer et al., 2014), we checked whether participants read and paid attention to our instructions. Table 3 shows the failure rates for all samples on the two attention-check questions employed in the study (one in the beginning and one in the middle). Whereas only 14% of MTurk participants failed both questions, almost half of the CBDR participants failed them, and the majority of the participants from all other sites failed them as well. Interestingly, CrowdFlower participants (who showed the fastest response rate) had a failure rate of almost 75%. Thus, from the perspective of participants' attention and compliance with written instructions, no site showed an adequate alternative to MTurk.

Table 3. Percent of participants who failed the attention check questions.

Sample	Failed first	Failed second	Failed both
MTurk	11.50%	3.50%	14.00%
CBDR	45.40%	5.20%	46.90%
CrowdFlower	72.90%	33.00%	74.40%
MicroWorkers	58.50%	19.70%	60.60%
RapidWorkers	93.30%	69.50%	93.30%

Reliability. We compared Cronbach's alpha measures for the three scales used in the study (IUIPC, NFC and RSES) across samples, and between participants who passed or did not pass both attention-check questions as reported above. As shown in Table 4, MTurk participants had the highest reliability scores on all three scales, followed by CrowdFlower participants, CBDR and Microworkers, all of whom performed adequately well on all scales (except for a somewhat lower score for CrowdFlower participants on the NFC scale). RapidWorkers participants showed high reliability on the IUIPC scale, but very low reliability on the NFC and mediocre reliability on the RSES scales. We used Hakistan & Whalen's (1976) method to compare between independent reliability coefficients and found no statistically significant

differences between the samples (using all participants from each sample) for the IUIPC, $\chi^2(4) = 6.63$, $p = .17$, but we did find statistically significant differences for the NFC and the RSES, $\chi^2(4) = 127.07, 75.69$, $p < .01$. As can be seen in Table 4, reliability measures for participants who passed both attention-check questions, compared to those who did not, were typically not different, and sometimes even in the opposite expected direction. Only for the NFC scale we found a higher reliability measure for those who passed compared to those who did not pass among the sample from MTurk, CrowdFlower and MicroWorkers. To summarize, from the perspective of reliability on established scales, MicroWorkers, followed by CrowdFlower, provided the best alternative to MTurk, whereas RapidWorkers results were less than satisfactory.

Table 4. Reliability (Cronbach's alpha) for the different samples.

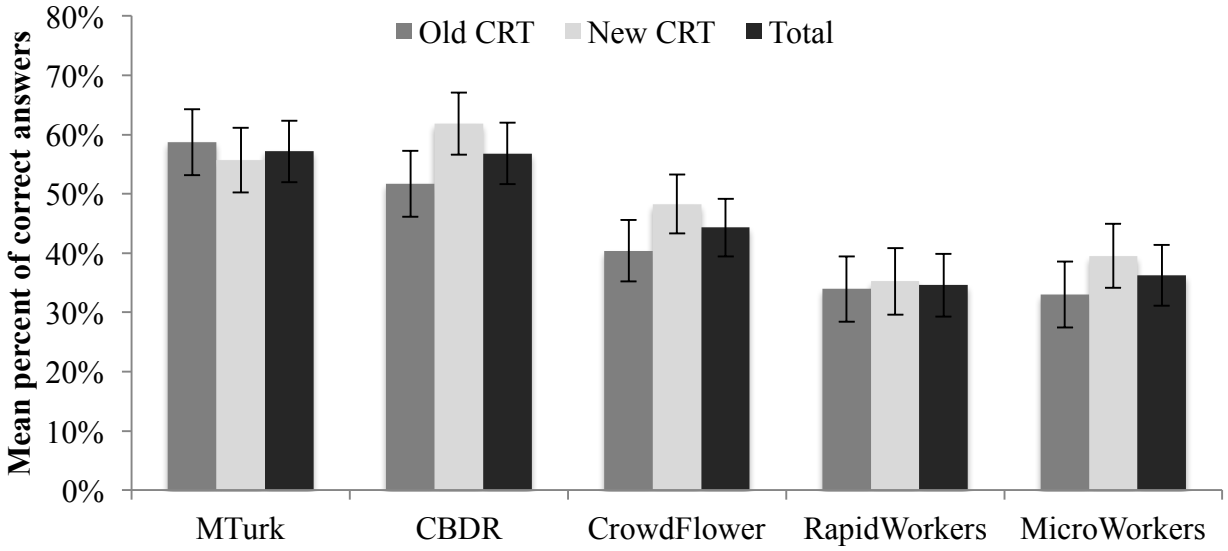
	IUIPC			NFC			RSES		
Passed both	No	Yes	All	No	Yes	All	No	Yes	All
MTurk	0.93	0.876	0.889	0.845	0.91*	0.901	0.873	0.931*	0.925
CBDR	0.891	0.873	0.883	0.828	0.895	0.871	0.899	0.92	0.911
CrowdFlower	0.875	0.901	0.887	0.607	0.771*	0.681	0.794	0.855	0.815
MicroWorkers	0.872	0.83	0.857	0.793	0.879*	0.834	0.874	0.908	0.891
RapidWorkers	0.847	[a]	0.84	0.155	[a]	0.354	0.661	[a]	0.689

* denotes a statistically significant difference ($p < .05$) compared to the participants in that sample who did not pass both questions. [a] Not calculated due to a small sample size ($n = 7$).

Non- naivety. We computed a percent score of correct answers for the old and new versions of the CRT separately, as well as a combined total score. As seen in Figure 2, MTurk and CBDR participants obtained the highest scores overall, followed by CrowdFlower,

RapidWorkers and MicroWorkers. The differences across samples on all three scores were statistically significant, F (4. 889) = 15.84, 14.53, 15.76, $p < .01$, respectively. We examined pair-wise differences on the total score and found, after Bonferroni correction, significant differences when comparing MTurk and CBDR to the other samples ($p < .01$), but no differences between MTurk and CBDR ($p = 1$). On average, MTurk and CBDR participants outperformed the other samples' participants by 17.76% of correct answers (56.89 vs. 39.21, SD = 36.9, 34.04, respectively, t (888) = 7.44, $p < .01$, d = 0.50). No statistically significant differences were found between the other three samples. Thus, it appears that CrowdFlower, RapidWorkers, and MicroWorkers contain more naïve participants compared to MTurk (and presumably CBDR).

Figure 2. Mean percent of correct answers (and 95% CI) for the old CRT, the new CRT and a total score (average of both) in the different samples.



Reproducibility. We next examined effect sizes on bona-fide effects from the judgment and decision-making literature: the Asian-disease gain vs. loss framing, the sunk-cost fallacy, and anchoring-and-adjustment (four items). Table 4 shows the effect sizes for these tasks across

sites. For the gain vs. loss framing effect, all but RapidWorkers' participants showed the effect in the expected direction (over-choice of sure option in the positive vs. the negative framing), and these effects were statistically significant. However, only MicroWorkers and CBDR participants showed a statistically significant effect in the sunk-cost task (although the effect was still in the expected direction on MTurk and CrowdFlower). For the anchoring tasks, we first computed standardized scores for participants' estimations, and then conducted a MANOVA on the four estimations with anchoring condition and sample as between-participants factors. We found a statistically significant effect for both the anchoring condition and the sample, as well as their interaction, Wilk's λ = .78, .90, .89, F (4/16, 326/996) = 22.77, 2.24, 2.43, p < .01, respectively. Subsequent ANOVA and post-hoc comparisons (using Bonferroni's correction) found statistically significant effects between anchoring conditions only for the Mississippi and Chicago questions, F (1, 329) = 79.75, 13.40, p < .01; an effect for sample only for the Mississippi question, F (4, 329) = 4.61, p < .01; and an interaction effect only for the Mississippi question, F (4, 329) = 4.58, p < .01. As can be seen in Table 5, the effects of the anchoring was not different between samples, except for an unusually high effect for the Mississippi question in the RapidWorkers sample. To summarize, from the perspective of replicability of known effects, both CrowdFlower and MicroWorkers could be considered a good alternative for MTurk, whereas RapidWorker's results were less than adequate.

Table 5. Effect sizes across samples (p-values are in parantheses).

	Asian disease	Sunk Cost	Anchoring Cohen's *d*			
	Chi-square	Cohen's *d*	*M*	*C*	*E*	*B*
MTurk	26.12 (<.01)	0.16 (0.25)	0.71	0.42	0.52	-0.24

			(<.01)	(0.06)	(0.02)	(0.28)
CBDR	17.69 (<.01)	0.49 (<.01)	1.0 (<.01)	0.41 (0.07)	-0.29 (0.18)	0.19 (0.38)
CrowdFlower	27.77 (<.01)	0.22 (0.12)	1.16 (<.01)	0.3 (0.07)	0.86 (0.18)	0.29 (0.38)
MicroWorkers	23.67 (<.01)	0.54 (<.01)	0.69 (<.01)	0.43 (0.1)	-0.26 (0.31)	-0.25 (0.32)
RapidWorkers	2.44 (0.12)	-0.02 (0.91)	5.86 (<.01)	0.79 (0.04)	2.23 (<.01)	0.43 (0.23)

Note: For the anchoring tasks, *M* = length of the Mississippi; *C* = population of Chicago; *E* = height of the Everest; *B* = number of babies born in the U.S.

Dishonest behavior. Comparing the distribution of die outcomes to a uniform random distribution showed no statistical differences among any of the samples that received this part of the study (MTurk, CBDR and MicroWorkers; p = .83, .13, .51, respectively), showing no evidence for dishonest reporting on the die roll task. Thus, from the perspective of participants' honesty, MicroWorkers cannot be ruled out as an alternative to MTurk. We found this result – practically no cheating in any of the sample – quite surprising given the relevant literature, and we discuss several possible (albeit post-hoc) explanations for it in the Discussion section.

To summarize thus far, it appears that both CrowdFlower and MicroWorkers exhibit a high level of data quality, not inferior to that of MTurk's on most respects, whereas RapidWorker's results were unsatisfactory on most of the aspects examined in this study. The following section presents data not directly related to data quality, but that could still be of interest when selecting a platform for online research.

Overlap of participants between sites. We asked participants from each platform to indicate the frequency with which they have used that platform, as well as the other crowdsourcing platforms (not including CBDR) by stating whether they use it "all the time,"

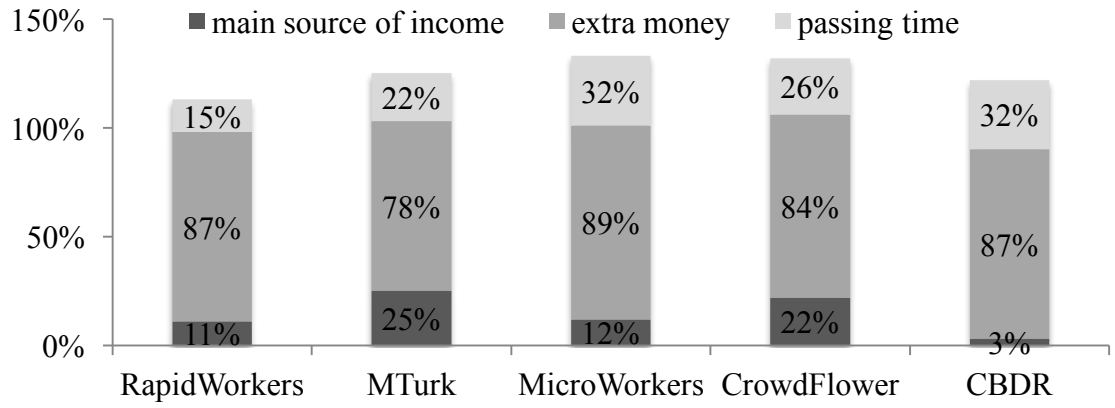
"many times," "a few times," "once or twice," or "never." We then grouped the percentages of participants that reported using the different platforms and found, generally, low degrees of overlap of participants between the different platforms, as shown in Table 6. MTurk and CrowdFlower participants seem to be almost solely focused on using their respective platforms, and CBDR participants rarely used any of these platforms. In contrast, RapidWorkers participants sometimes used MicroWorkers and CrowdFlower (but not MTurk), and MicroWorkers' participants sometimes used MTurk.

Table 6. Percent of participants reporting using the different platforms "many times" or "a lot of times" between the different samples.

	Uses MTurk	Uses CF	Uses MW	Uses RW
MTurk	93.50%	3.00%	1.50%	1.50%
CrowdFlower (CF)	4.93%	91.13%	2.96%	2.96%
MicroWorkers (MW)	24.47%	5.85%	72.87%	3.19%
RapidWorkers (RW)	8.57%	10.48%	29.52%	73.33%
CBDR	3.09%	0.52%	0.52%	0.52%

Reasons for using the site. As Figure 3 shows, most of the participants from all of the platforms stated that they use the platform mainly to earn some extra income. However, a fair amount of participants from MTurk and CrowdFlower (25% and 22%) stated that the platform is their main source of income. Many MicroWorkers and CBDR participants indicated they use the platform for passing time.

Figure 3. Distribution of reasons for using the platform (multiple-choice question).



Individual differences between platforms. We examined whether the sampled participants differed between the sites on several dimensions: education, income, ethnicity, country of residence, and U.S. citizenship. We found significant differences in ethnicity (χ^2 (4) = 233.11, p < .01), with MTurk, MicroWorkers and CBDR participants being predominantly Caucasian (more than 60%), whereas CrowdFlower having a relatively high rate of Asian participants, as did RapidWorkers and CBDR. Interestingly, over a quarter of the participants from RapidWorkers actively refused to divulge their ethnicity by marking "I prefer not to say" (see Figure 4). The distribution of country of residence also differed significantly between the sites (χ^2 (12) = 688.53, p < .01): whereas the majority of users from all sites except CrowdFlower came from the U.S. (see Figure 5), about half of the users on CrowdFlower were from Europe, and the rest were from India (16.3%) or other (mostly South American and Asian) countries (28.1%). These differences clearly show that CrowdFlower taps into completely different populations than MTurk, namely European participants that are absent on MTurk and the other sites. We discuss the implications of these differences in the Discussion section.

We also found differences in education level, income level and U.S. citizenship between samples (χ^2 (4= 128.37, 125.84, 568.74, p < .01): MTurk, CBDR and CrowdFlower participants

mostly held an undergraduate degree, whereas MicroWorkers participants mostly had a high-school education level. The median income among MTurk and MicroWorkers participants was slightly higher ($25K-$50K) than among the other sites ($10K - $25K). More than 80% of the participants on all sites except CrowdFlower were U.S. citizens or permanent residents, compared to only 4.9% on CrowdFlower (unsurprisingly). Again, we found high degrees of non-disclosure mostly among RapidWorkers participants as 30% of them refused to divulge their income, and 24% refused to say whether they are U.S. citizens or not (compared to less than 10% among all other samples, except one case - 22% of CBDR participants refused to divulge their income).

Figure 4. Distribution of ethnicity across samples.

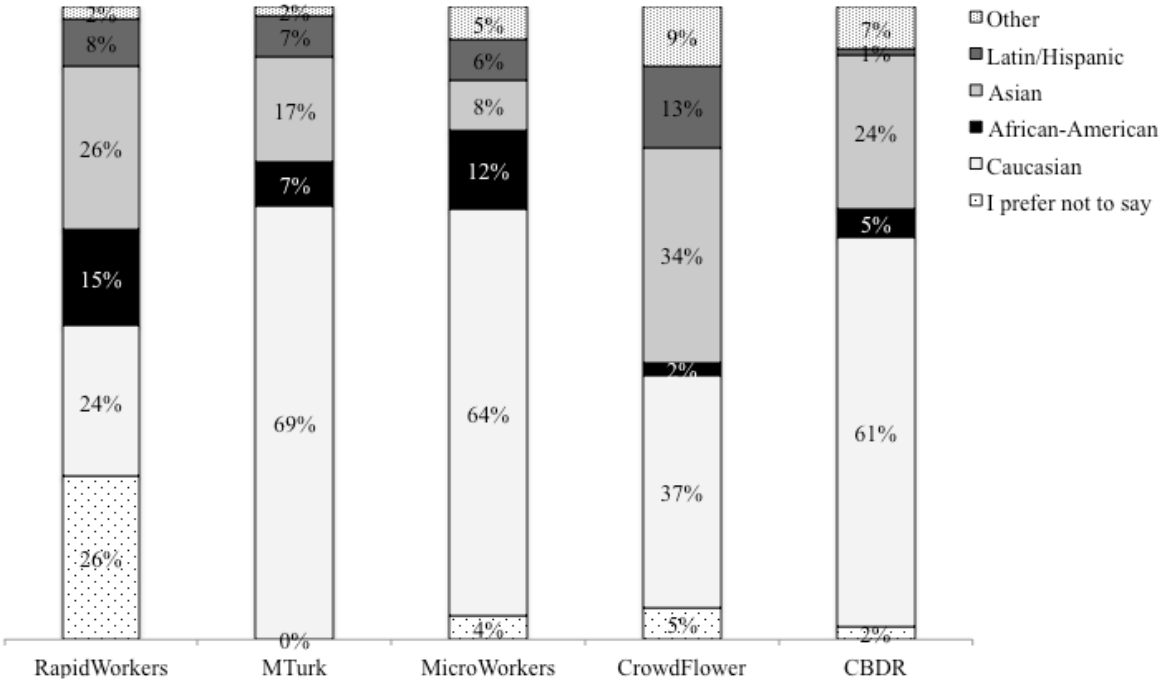
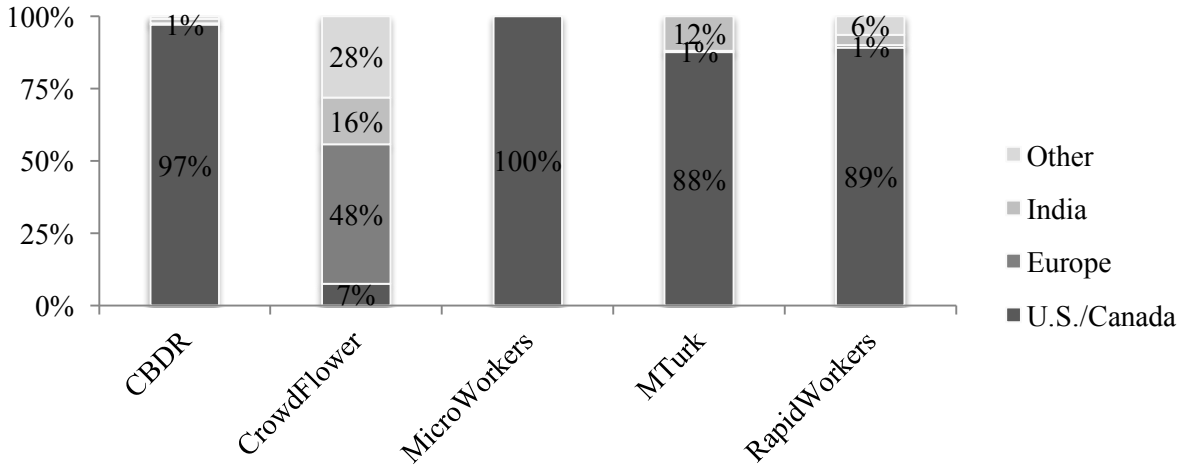


Figure 5. Distribution of country of residence across samples.



To summarize, it appears that CrowdFlower is distinctly different from the other samples in aspects of ethnicity, origin, but not in aspects of income and education (in which MicroWorkers participants were a little bit less educated, but of higher income, than the other sites). Another important and interesting difference found between the sites was RapidWorkers participants' general reluctance to disclose personal information as about a quarter of them actively refused to divulge details such as income, ethnicity, or U.S. citizenship. We discuss the possible implications of these differences in the next section.

Discussion

Our empirical investigation of the selected platforms suggests that among the three viable alternatives on which we could conduct an online survey and obtain responses, RapidWorkers did not provide adequate results and researchers should be cautioned against using it at present time. In contrast, both CrowdFlower and MicroWorkers seem to be possible alternatives for MTurk: The reliability of questionnaire on those samples was adequate on most cases; the samples typically reproduced the known effects of the Asian-disease problem, sunk-cost and anchoring; and MicroWorkers participants did not try to cheat the researcher by over-reporting their performance. Comparing between these two platforms, CrowdFlower showed superiority

over MicroWorkers in terms of response rates (achieving a response rate identical to that of MTurk's). However, about 75% of participants on CrowdFlower (compared to about 60% on MicroWorkers) failed both the attention-check questions, suggesting these participants may not fully read the instructions in the survey questions. This did not, however, seem to impact their ability to provide reliable responses to the questionnaires or reproduce the effects in the known tasks. Additionally, the high rates of passing attention-check questions on MTurk could be due to participants' past experience with these, or similar, attention-check questions (Chandler et al., 2014; Peer et al., 2013), and a high failure rate could actually be considered desirable because it implies these sites' participants are still somewhat naïve to experimental materials. A summary comparison of the differences between the sites is given in Table 7.

Table 7. Summary of differences between the sites.

	MTurk	CrowdFlower	MicroWorkers	RapidWorkers	CBDR
Dropout rate	Low	Low	High	Low	Low
Response rate	Fast	Fast	Medium	Slow	Medium
Attention-check failure rate	Low	High	High	High	High
Reliability	High	High	High	Low	High
Reproducibility	Good	Good	Good	Poor	Good
Naivety	Low	High	High	High	Low
Honesty	High	-	High	-	High
Ethnic diversity	Low	High	Low	Low	Low
Geographic origin	Mostly U.S.	Mainly Europe	Mostly U.S.	Mostly U.S.	Mostly U.S.
Active non-disclosure	Low	Low	Low	High	Low

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., & Feng Kao, C. (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. (In press). Non-naïve participants can reduce effect sizes, *Psychological Science.*

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.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 25-42.

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.

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

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*, 1161-1166.

Litman, L., Robinson, J., & Rosenzweig, C. (2014). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, 1-10.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. *Information Systems Research*, *15*(4), 336-355.

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

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.

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