Beyond the Turk:

Alternative platforms for crowdsourcing behavioral research

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

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

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 exhibits slowing rates of population replenishment, and growing participants' non-naivety. Recently, a number of alternative platforms have emerged, offering capabilities similar to MTurk while providing access to new and more naïve populations. We examined two such platforms, CrowdFlower (CF) and Prolific Academic (ProA). We found that both platforms' participants were more naïve and less dishonest compared to MTurk. CF showed 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. We also found important demographic differences between the platforms. We discuss how researchers can use these findings to better plan online research, and their implications for the study of crowdsourcing research platforms.

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) as an efficient crowdsourcing platform for the recruitment of online human subjects for research (Paolacci, & Chandler, 2014). A large body of work has shown that MTurk is a reliable and cost-effective source of high-quality and representative data, for various research purposes, in and outside of behavioral sciences (e.g., Buhrmester, Kwang, & Gosling, 2011; Chandler, Mueller, & Paolacci, 2014; Crump, McDonnell, & Gureckis, 2013; Fort, Adda, & Cohen, 2011; Goodman, Cryder, & Cheema, 2013; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010; Peer, Vosgerau, & Acquisti, 2013; Rand, 2012; Simcox, & Fiez, 2014; Sprouse, 2011). However, a current concern associated with using 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,” completing common experimental tasks and questionnaires more than once. This high rate of non-naivety among MTurk participants has been recently shown to have the potential to significantly reduce effect sizes of known research findings (Chandler et al., 2015). This, combined with the recent findings that an average lab samples from an effective population size of only around 7,000 participants, and that this population also has a slow replenishment rate (Stewart et al., 2015), suggests that alternatives should be explored.

Recently, several alternative platforms have emerged, offering services that are similar to MTurk. These alternative platforms offer access to new and 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). Access to alternative crowdsourcing platforms for recruiting human subjects could be highly beneficial for researchers interested in conducting online surveys and experiments, as long as these new platforms provide high-quality data. In this paper, we compare the data quality of selected alternative platforms with data collected both via MTurk and a university-based online participant pool. We compare these alternative services along several dimensions essential for online behavioral research, including response rates, attention, dishonesty, reliability, and the replicability of existing research findings.

Table 1. Comparison of platforms' properties and features

	Mturk	**CF**	**ProA**	**CBDR**
Population size[a]	Over 500K	Over 10K	About 20K	About 10K
Screen by reputation	Yes, built-in	No option	Yes, built-in	No option
Screen Ps by demographics	By using qualifications	By location and language only	Yes, built-in	Yes, built-in
Exclude past participants	By using qualifications	No option	Yes, built-in	Yes, built-in
Automatic approval	After preset time	On survey completion	On survey completion	After preset time
Bonus method	Individual or in bulk (requires scripting)	Individual	Individual or in bulk	None in the system

[a] As reported by the website administration.

After searching for and testing several available crowdsourcing websites, we identified and focused on two selected platforms similar to Mechanical Turk in design and purpose:

CrowdFlower (CF) and Prolific Academic (ProA)[1]. Table 1 summarizes some key properties and features that differ between these platforms. We also included, as a comparison group, participants from the Center for Behavioral Decision Research (CBDR) participant pool, a more traditional pool that includes student and non-student participants, and is managed out of Carnegie Mellon University. Many research institutions have access to participant pools of their own. They may differ from the CBDR pool, but there may also be many commonalities, including composition and retribution models. There is therefore much one can learn from by sampling from a such pool.

Method

Sampling and participants. We sampled 200 participants from each of four platforms: CF, ProA, CBDR, and MTurk. 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 reach the goal of recruiting approximately 200 participants from all platforms, and ended 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, mainly because we wanted to assess differences between the platforms on these aspects. Participants on MTurk and CF were paid $1

[1] In addition to CF and ProA, we also examined MicroWorkers, RapidWorkers, MiniJobz, ClickWorker and ShortTask. These websites were not as effective as the ones we have reported on in this paper, either in their data quality or response rate or the cost of recruitment, and so we do not discuss them in this paper.

for survey completion, participants on ProA received £1, and participants on CBDR received a chance to win a $50 gift card given to one out of every 50 participants.

Procedure. The survey included several stages. The first stage consisted of several questionnaires and experimental tasks adopted from prominent studies in psychology, which were used to test data quality. 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 that tested dishonest behavior.

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

Sample	Started the study	Completed	Percent of dropouts	Percent males	Mean age (SD)
MTurk	220	201	8.6%	56.7%	33.7 (9.4)
CF	238	221	7.1%	73.6%	32.2 (9.6)
ProA	243	214	11.9%	64.5%	31.3 (11.3)
CBDR	215	195	9.3%	29.2%	28.9 (13.5)

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 et al., 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 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, 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 6’s [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?” 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, the participants then answered demographic questions, and questions that pertained to the use of their respective platform and other platforms (as detailed in the Results section). The next and 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 down. 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, this task gave participants 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

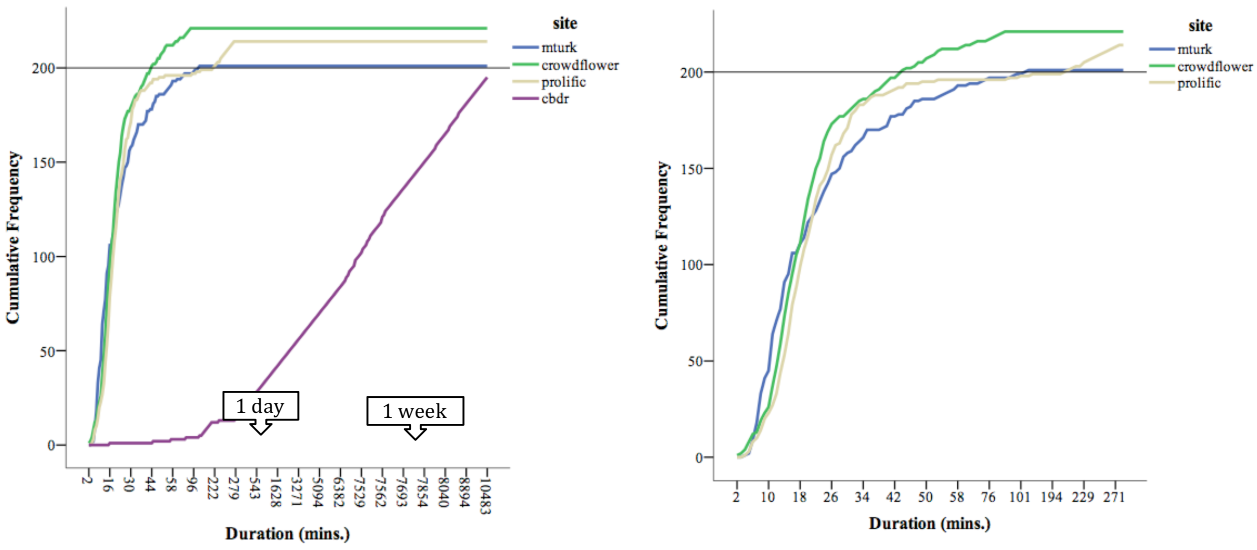
Response rates. As detailed in Table 2, dropout rates ranged around 10% for all platforms, with no significant differences between the platforms, χ^2 (3) = 3.43, *p* = .33. All of the following analyses include only participants who completed the entire study.

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 end 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, which took 1:48 hours to collect 200 responses. ProA took 4:37 hours to produce 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 rates, while ProA had a somewhat slower response rate overall than these two 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, as illustrated by Figure 1b. 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 1a and 1b. Response rates across all samples (1a) and without CBDR (1b).



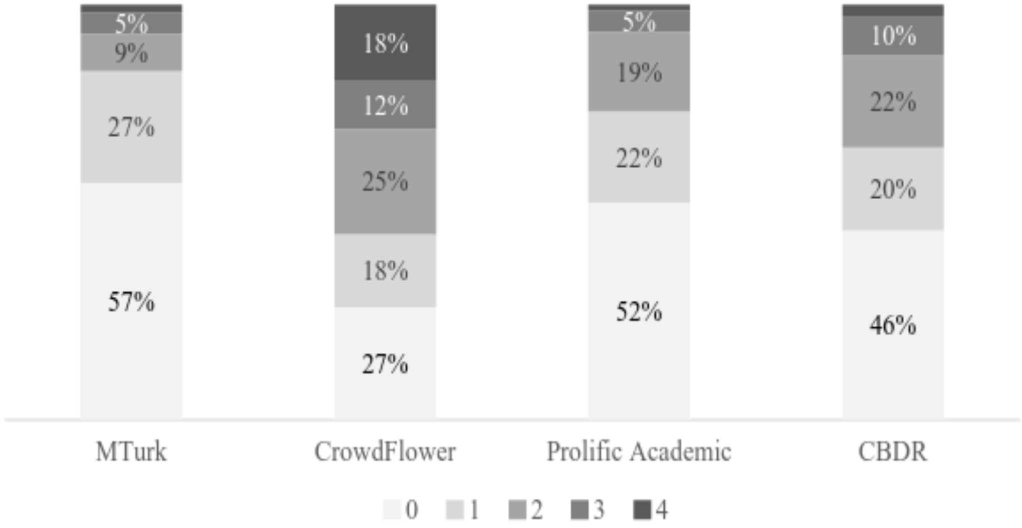
(1a)

(1b)

Attention. Using the four attention-check questions, we tested whether participants read and paid attention to our instructions. Figure 2 shows that more than half of MTurk's and ProA's participants passed all ACQs, whereas only 27% of CF's participants passed all ACQs and 18.1% of CF's participants failed all ACQs. CBDR participants performed better than CF, with 45.6% of participants passing all ACQs. These differences were statistically significant, χ^2 (3) = 122.37, $p < .01$. 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

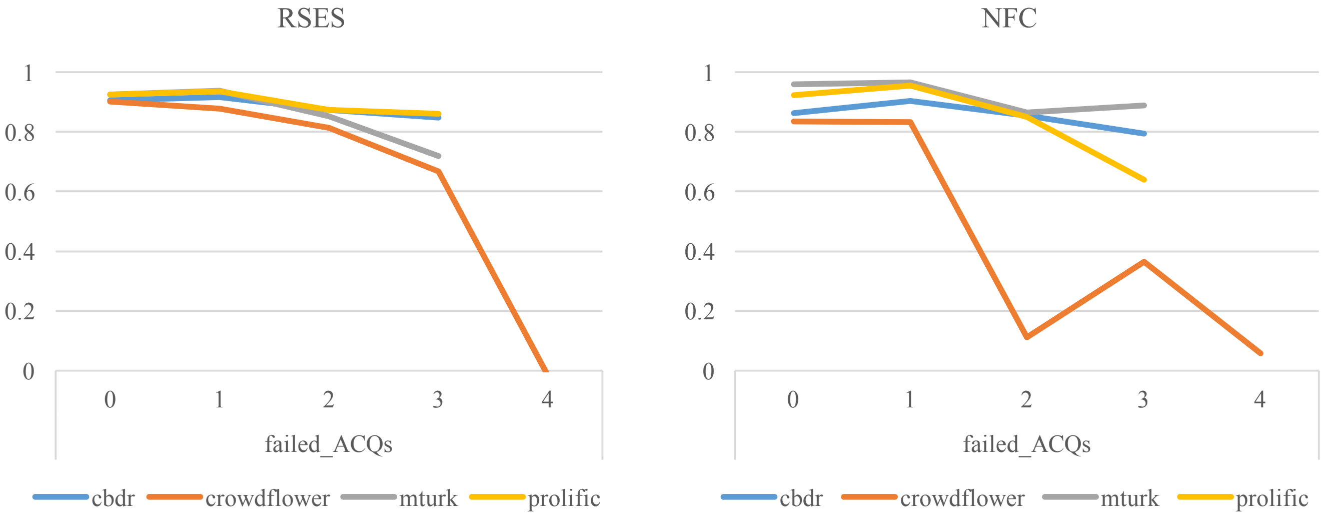
ACQs (SD=0.96), ProA participants failed 0.81 ACQs (SD=1.01), CBDR participants 1.04 ACQs (SD=1.14) and CF participants failed the most: 1.76 ACQs on average (SD=1.44). All post-hoc differences, except between ProA and CBDR, were statistically significant, after applying Bonferroni's correction (p < .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.

Figure 2. Percentage of participants according to number of failed attention check questions between the platforms.



Reliability. We compared internal reliability measures (Cronbach's alpha) for the RSES and NFC scales used in the study between samples, and as a function of the number of ACQs failed by participants (following Peer et al., 2013). 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 (above 0.80) for all samples, even for participants who failed one or two ACQs. However, MTurk and CF participants who failed three ACQs showed lower results (alpha = 0.718 and 0.668, respectively), and for the 40 participants who failed all ACQs in the CF sample, reliability was null (alpha = -0.018). A somewhat similar pattern emerged with the NFC scales (Figure 3b), except that the decline in reliability amongst CF participants started at the 2nd failed ACQ. Among participants who failed three ACQs, CF's participants exhibited the lowest values (alpha = 0.35), as compared to ProA (0.64), CBDR (0.794) and MTurk (0.889). Using Hakstian and Whalen's (1976) method to compare between independent reliability coefficients, we found these specific differences to be statistically significant ($p < .05$).

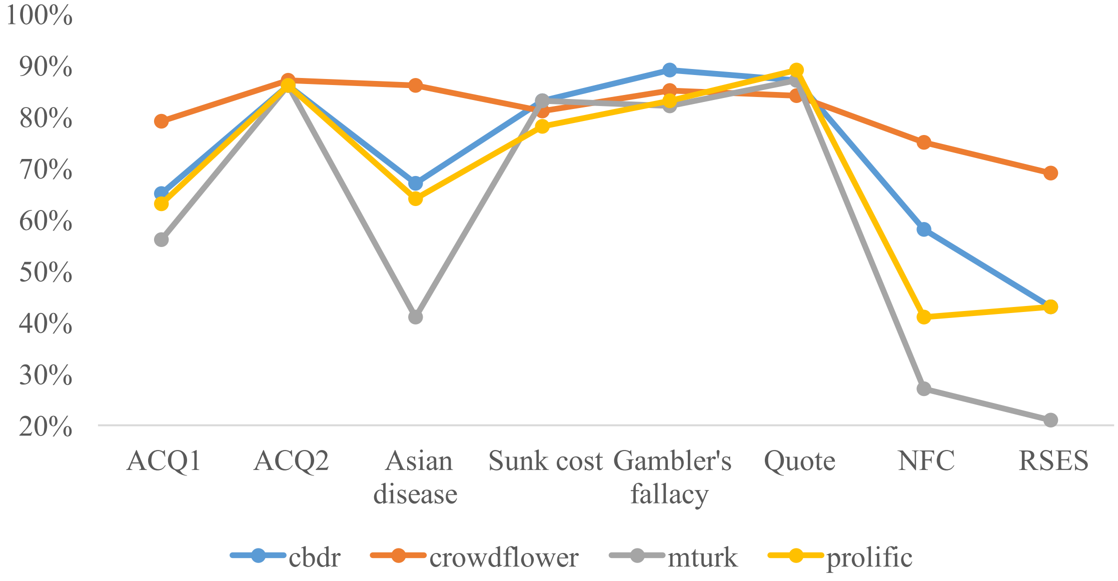
Figure 3a and 3b. Cronbach's alpha for the RSES (3a) and NFC (3b) between the platforms and as a function of the number of failed ACQs.



Note: Values of sub-groups that had less than 10 participants are not displayed.

Non-naivety. As mentioned earlier, participants were asked, after each experimental task, questionnaire, and the first two ACQs, whether that had been the first time 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.

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



Reliability of all eight tasks' dichotomous scores of familiarity was adequately high (alpha = 0.744), so we computed the percentage of tasks that 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 a mean percent 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.

Reproducibility. We next examined the effect sizes of the four experimental tasks used in the study. As Table 3 shows, we were able to replicate all effects in MTurk and ProA samples, but CF participants did not exhibit either the Sunk Cost or Gambler’s Fallacy effects. CBDR participants did not exhibit the Gambler’s Fallacy effect either.

Table 3. Effect sizes (Cohen’s d) across samples

	Asian disease	Sunk Cost	Gambler’s Fallacy*	Quote attribution
MTurk	**0.82**	**0.27**	**0.28**	**0.73**
CBDR	**0.76**	**0.42**	0.12	**0.51**
CF	**0.72**	0.02	0.20	**0.54**
ProA	**0.63**	**0.39**	**0.29**	**0.68**
Overall	**0.72**	**0.27**	**0.22**	**0.61**

Note: bold values denote statistically significant differences between the conditions, p < .05.

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

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 with regard to each individual participant whether they 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, respectively, *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).

To summarize the comparison of data quality properties between the platforms, we found that, compared to MTurk, CF showed the 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's participants reported less familiarity (higher naivety) regarding common experimental tasks, two of these tasks could not be replicated on that sample, whereas all tasks replicated on ProA, even though ProA participants reported higher naivety than MTurk participants. Lastly, both ProA and CF participants showed lower degrees of dishonest behavior compared to MTurk. We elaborate on the potential factors underlying these differences, and their implications for online behavioral research, in the Discussion section.

Ethnicity, location and language. We found statistically significant differences in the distribution of ethnicity between the samples, χ^2 (15) = 92.64, p < .01. As Figure 5 shows, while Caucasians represented the majority of participants on MTurk and ProA, they were represented less on CBDR and, especially, on CF, which included a higher proportion of Asian participants (around a quarter). CF also included a substantial proportion (16%) of Latin/Hispanic participants, who represented only 4% of the other samples. Figure 6 shows the distribution of reported location[2] between the platforms. While MTurk's (and CBDR's) vast majority of participants came from North America (U.S. and Canada), CF and ProA showed 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%) or South America (4%). In CF, in contrast, only 5% came from North America, with the majority of participants from Europe (43%), and another 25% of participants from East Asia or India. The vast majority of participants on MTurk, ProA, and CBDR reported that they could read English at a "very good" or "excellent" level (99%, 97.2%, 91.8%, respectively), versus only 69.2% among CF participants (the rest rated their reading ability as "good" or worse).

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

Figure 5. Ethnicity distributions between the platforms.

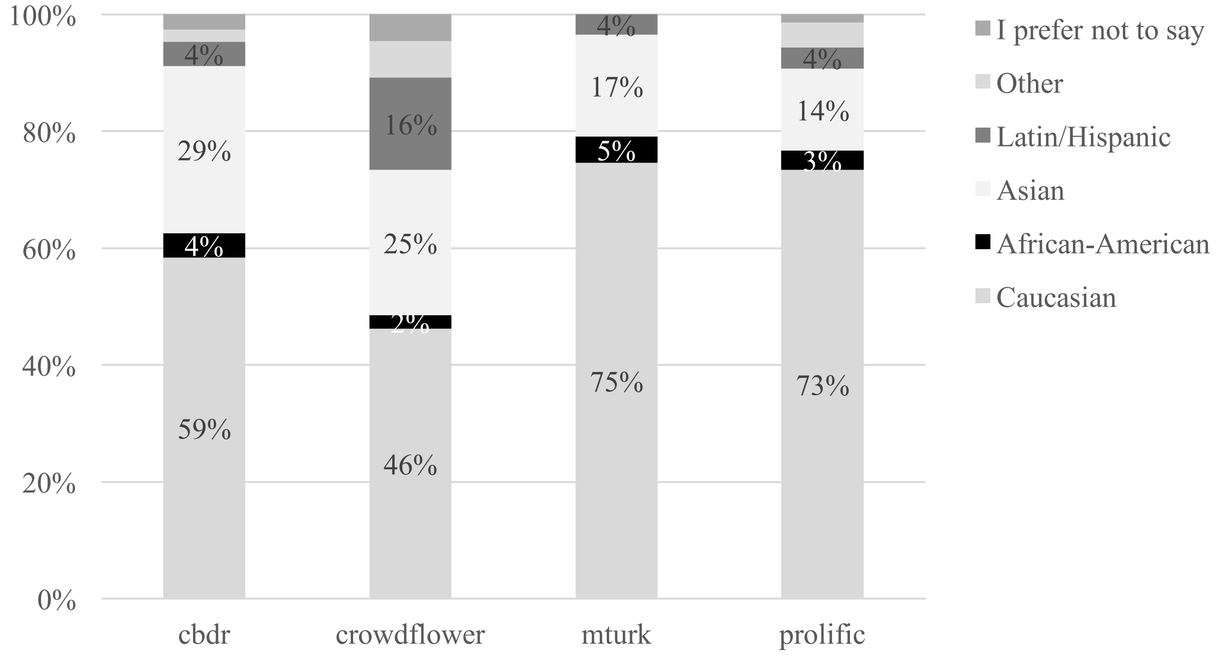
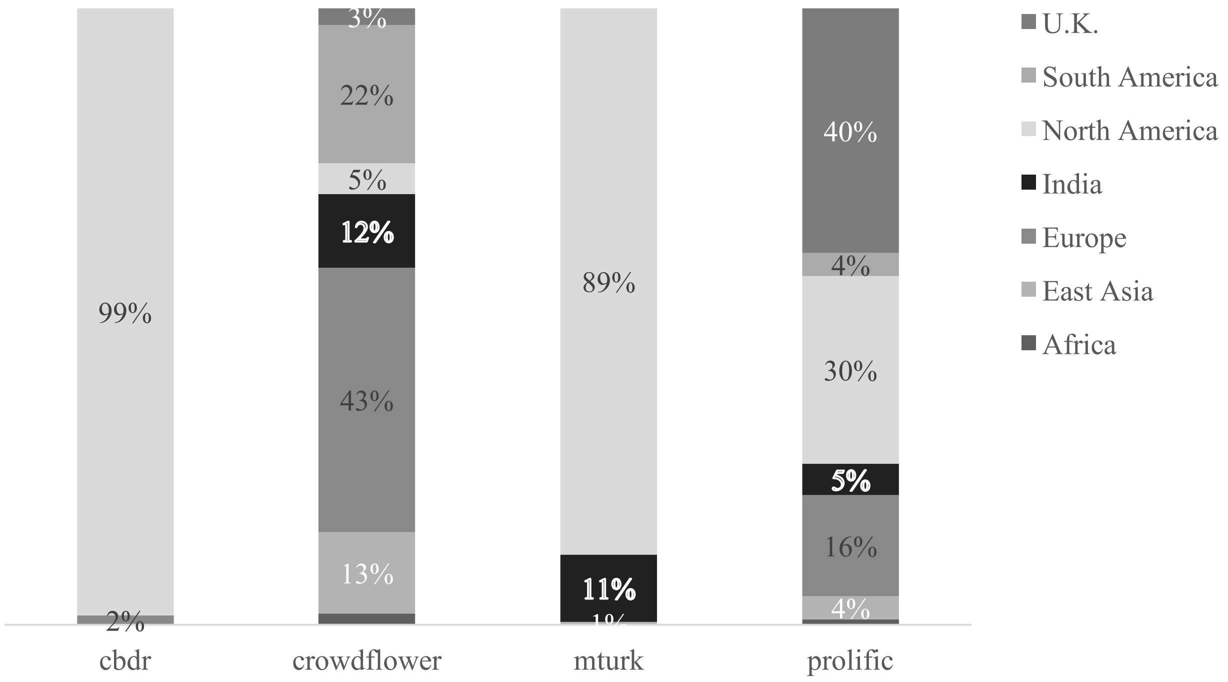
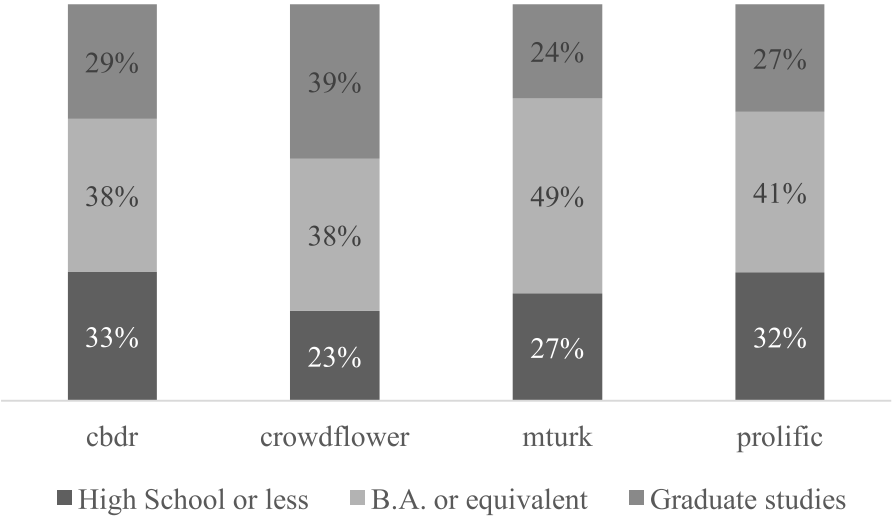


Figure 6. Reported location distributions between the platforms.



Education and income. We found statistically significant differences in the distribution of education, χ^2 (6) = 17.85, $p < .01$. CF participants seemed to be highly educated, with the highest percentage of participants reporting some degree of graduate studies and the lowest percent of participants reporting high-school (or lower level) education. Regarding income, we found that the median income on MTurk was between $25K-$50K whereas for all the other samples it was $10K-$25K. A Kruskal-Wallis test showed this difference was significant, $p < .01$.

Figure 7. Education distributions between the platforms.



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 use 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, most MTurk participants report spending between 8 and 40 hours per week on the platform, whereas most CF participants spend more, between 20 and more than 50 hours per week. ProA users spend considerably less time, with most reporting spending between 1 and 4 hours per week only (CBDR users report spending even less). As Figure 9 shows, this clearly results in differences in earnings for participants between the platforms: whereas more than 70% of MTurk-ers report 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, which could be due to students receiving academic credit instead of money). 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.

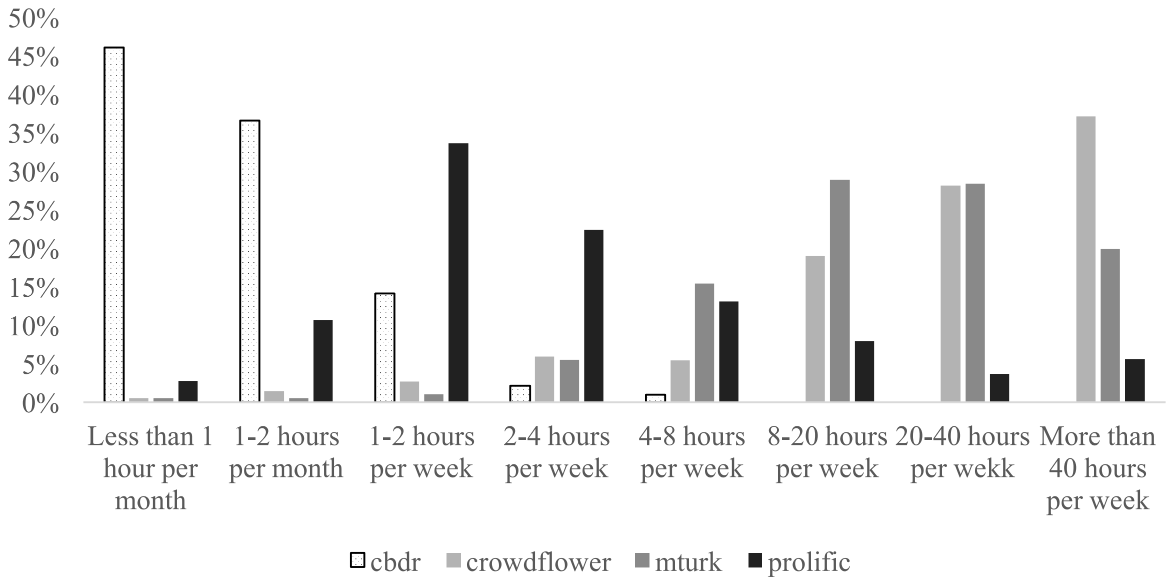
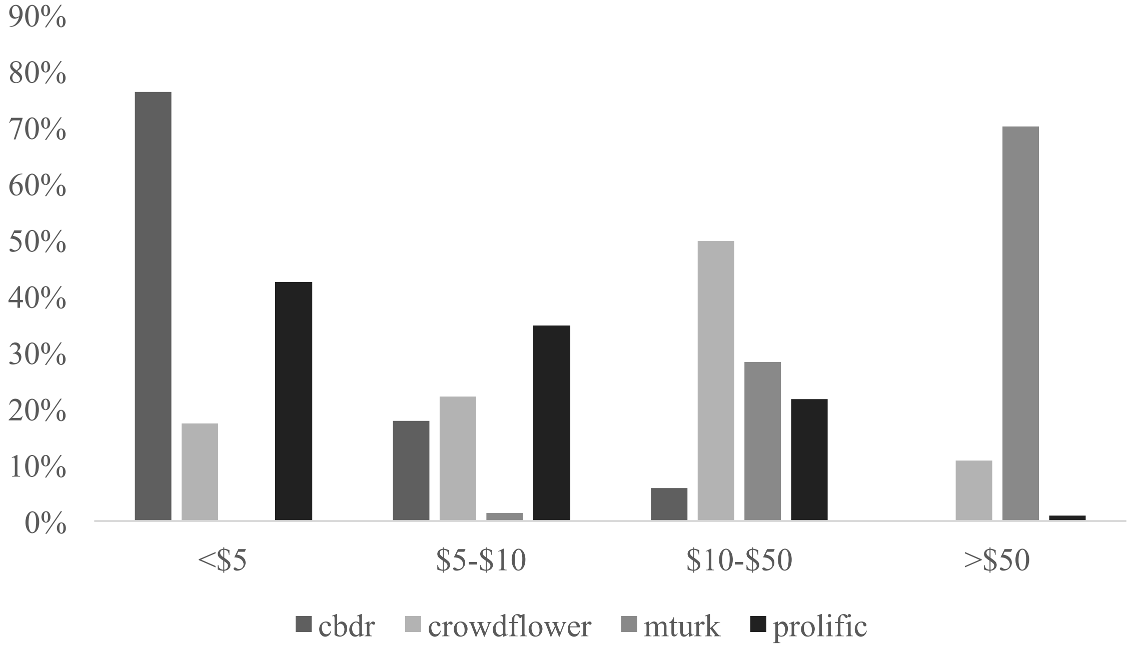


Figure 9. Percentage of participants in different quartiles of average weekly earning between the platforms (the cutoffs represent the quartiles of earnings in the overall sample).



Discussion

Our empirical investigation of various selected platforms suggests that while both CF and ProA show adequate data quality, ProA seems to be the most viable alternative candidate 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 behaviors, 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.

Interestingly, data reliability was significantly higher for ProA, as compared to CF, when participants failed attention checks. For participants who failed all attention checks, we actually observed null reliability on the RSES for CF; for those who failed two checks, we observed close to null reliability on the NFC scale. That is not the case for ProA, where reliability only shows a slight decrease as a function of the number of checks failed. 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 et al., 2014; Peer et al., 2013), and a high failure rate could actually be considered desirable because it implies naivety 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: both of these platforms provided participants with a lower propensity towards cheating, as compared to MTurk. A summary comparison of the differences found between the platforms is given in Table 4.

Table 4. 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	Highest	Low	Medium
Geographic origin	Mostly U.S.	Mainly Europe	Mostly U.S.	Mostly U.S.
English fluency	High	Lower	High	High
Income level	Low	Low	High	Low
Education level				
Usage frequency	High	Highest	Medium	Lowest
Typical tasks				
Average pay/week				
Overlap with other	Some (PA)	Few	Some (Mturk)	Few

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 considered if the study is to be run on multiple platforms, or if (for example) a similar study has already been conducted on MTurk. The other issue to consider is the demographic composition of these platforms. The most salient difference lies in participants' ethnicity and country. 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 majority of CF participants reside outside the U.S. (mainly in Europe and Asia), while both ProA and MTurk attract mostly U.S. residents. This suggests that the two platforms (CF and ProA) tap into two very 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 samples, as MTurk almost exclusively offers. In that sense, researchers might choose to take advantage of CF's (and to a lesser extent, ProA's) access to non-U.S. populations. In doing so, researchers might 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).

Interestingly, we found significant differences in (dis)honesty between MTurk and the other platforms, as participants from all the tested alternatives seemed to report their die roll honestly at higher rates. 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 being for academic research, MTurk's appeal is more about earning money quickly.

The results of the current research could guide researchers' choices when venturing with online crowdsourcing research, but additional research should be conducted to explore some unanswered questions that stem from the current study's limitations. First, the roots and causes of the differences found between the platforms remain obscure, as we could only control the sampling (and not allocation) of participants from the different samples. Second, it remains an open question how constant or transient any of the findings actually are. While some differences could be considered, presumably, as relatively stable (e.g., response rates; demographics; propensity for dishonest behavior), many others (e.g., attention, naivety, etc.) could be much more temporary. In this regard, the current paper offers a helpful framework by which platforms could be evaluated overtime (and also following certain events, such as a major update in pricing). This framework, that includes measures of attention, reliability, reproducibility, naivety and dishonesty, could also be used to evaluate new platforms, such as may arise in the future. Such studies would ensure that the goal of the current study – to empirically test and identify viable and effective platforms for online research – would not be forgone.

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.

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.

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 – Description of the attention-check questions (ACQs) used in the study

The first ACQ 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 but to select "two" for the first question, add three to that number and use the result to answer the second question. Any response, other than "2" and "5" respectively, was coded as failing this ACQ.

The second ACQ was a novel question we designed for the purposes of this study. Participants were asked to follow these instructions: "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." However, the next paragraph actually instructed participants to only report zero: "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." In the next page, an image with five persons appeared, alongside a timer counting down from 20 seconds, and participants were asked to choose a number between 0 and 10. Any response other than zero was coded as failing this ACQ.

For the third and fourth ACQs, the item “I have never used the Internet myself” was added to the RSES, and the item “I currently don't pay attention to the questions I'm being asked in the survey” to the NFC scale. Any response other than “strongly disagree” was coded as failing these ACQs.