



Data quality of platforms and panels for online behavioral research

Peer Eyal¹ · Rothschild David² · Gordon Andrew³ · Evernden Zak³ · Damer Ekaterina³

Accepted: 23 August 2021

© The Psychonomic Society, Inc. 2021

Abstract

We examine key aspects of data quality for online behavioral research between selected platforms (Amazon Mechanical Turk, CloudResearch, and Prolific) and panels (Qualtrics and Dynata). To identify the key aspects of data quality, we first engaged with the behavioral research community to discover which aspects are most critical to researchers and found that these include attention, comprehension, honesty, and reliability. We then explored differences in these data quality aspects in two studies ($N \sim 4000$), with or without data quality filters (approval ratings). We found considerable differences between the sites, especially in comprehension, attention, and dishonesty. In Study 1 (without filters), we found that only Prolific provided high data quality on all measures. In Study 2 (with filters), we found high data quality among CloudResearch and Prolific. MTurk showed alarmingly low data quality even with data quality filters. We also found that while reputation (approval rating) did not predict data quality, frequency and purpose of usage did, especially on MTurk: the lowest data quality came from MTurk participants who report using the site as their main source of income but spend few hours on it per week. We provide a framework for future investigation into the ever-changing nature of data quality in online research, and how the evolving set of platforms and panels performs on these key aspects.

Keywords Data quality · Online research · Attention · Comprehension · Reliability · Honesty · Amazon mechanical turk · Prolific

In the past decade, there has been a proliferation of online platforms for research (e.g., Buhrmester et al., 2018; Goodman & Paolacci, 2017). This growth was brought on by the relatively lower cost of developing online panels and platforms, coupled with increased demand due to the relative ease of running research online, global occurrences (e.g., COVID-19), and trends (e.g., remote working) that have reduced options to run studies on campuses or laboratories. For example, the number of both participants and researchers on one online platform (Prolific) has reportedly increased more than sevenfold in the last 3 years, and a large share of this increase occurred at the outbreak of the COVID-19 crisis, in early 2020 (see Appendix). There is reason to believe that similar growth has occurred on other platforms, and while this

growth is probably desirable, it makes the issue of data quality of online studies even more important than before.

Data quality consists of varying attributes that have different levels of importance for different disciplines. For example, personality researchers would be concerned about the reliability and validity of scales (e.g., Schmidt et al., 2003); experimentalists would require high attention to stimuli and would require that participants carefully read and comprehend instructions (e.g., Rand et al., 2012), and if incentives are used in the design, researchers would require that participants not try to cheat for higher gains. Regardless of the discipline or goals of the research, some weighted combination of these and potentially other attributes combines into a concept of data quality that is core to the validity of any behavioral research, both for academic publication and for any real-world applications.

Previous research on online data quality focused on different attributes of quality among different types of audiences (for a comprehensive review see Thomas & Clifford, 2017). These mostly included testing participants' ability to pass attention-check questions (e.g., Paolacci et al., 2010, see also Göritz et al., 2021), reliability of validated scales (e.g., Buhrmester et al., 2016), replicability of known effects such as the Asian disease problem (e.g., Chandler et al., 2014; Feitosa et al., 2015; Goodman et al., 2013), representativeness of the sample (e.g., Ipeirotis, 2010), and non-naivety of

✉ captured/presentation of author names are correct and also if affiliated correctly." →Peer Eyal
eyal.peer@mail.huji.ac.il

¹ Federmann School of Public Policy, The Hebrew University of Jerusalem, Jerusalem, Israel

² Microsoft Research, New York, NY, USA

³ Prolific Inc., Newark, CA 94560, USA

participants to common research stimuli (Chandler & Paolacci, 2017). Most of these studies focused on comparing Amazon's Mechanical Turk (MTurk) to either student or online samples of convenience (e.g., Behrend et al., 2011; Goodman et al., 2013); fewer have also made comparisons between MTurk to other, newer platforms such as Prolific (Peer et al., 2017) or CloudResearch's Prime Panels (Chandler et al., 2019). The general conclusion from these studies has been that online platforms provide, in most cases, satisfactory data quality, especially compared with previous standards such as convenience samples of students.

In parallel, other studies explored the quality of survey responses from online panels (Porter et al., 2019; Thompson & Pickett, 2019; Wang et al., 2015; Yeager et al., 2011). These studies tend to focus on the difference between online panels and historical approaches such as random-digit-dialing telephone surveys, on their aggregated survey responses for key questions. This makes sense because the main concern of the survey industry is the representativeness of the results, not the quality of the individual responses. For example, MacInnis et al. (2018) focused on how secondary demographics (such as marriage and citizenship) and behaviors (such as moving or giving blood) from opt-in online panels and probabilistic online and telephone respondents are correlated. Hillygus et al. (2014) lay out a litany of knowledge and attitudinal differences in survey responses from panels (where respondents frequently answer surveys) versus random respondents (where respondents rarely answer surveys). However, there is not much in this literature that explores the quality of participants' responses, especially on online platforms and panels.

Previous research has suggested that reputation (approval ratings—share of participants' submissions that were accepted by the researcher) can positively predict higher data quality, as high reputation was found to produce better data quality, to the point that even attention-check questions were unnecessary (Peer et al., 2014). However, other research has shown that non-naïve participants can reduce data quality (Chandler & Paolacci, 2017). Thus, if highly reputable respondents might also be less naïve, it is unclear how these different aspects might interact to affect data quality. Furthermore, there have been claims that some MTurk users are extremely active, to the point that they might be considered “professional survey-takers,” which could hamper data quality as well (Chandler et al., 2014; Keith et al., 2017). Indeed, previous research has suggested that sites where participants' frequency of usage is lower (e.g., Prolific) show higher data quality (Peer et al., 2017). Thus, it is important to also examine how reputation (e.g., approval ratings) and frequency of usage, as well as the purpose for using the site (e.g., as the main source of income or not), are associated with different aspects of data quality.

In this paper, we present a framework for exploring data quality in online research. First, we determine what aspects of data quality researchers (in this case, behavioral researchers)

value the most, and we then go on to explore how key online audiences compare on these aspects. We extend previous research by examining online audiences that are popular but have not been examined much, and we include aspects of data quality that have been less explored in the past. Additionally, we consider the degree to which data quality could be predicted by participants' characteristics on the site.

Key data quality aspects for behavioral research

In order to inform our decision regarding the data quality aspects on which to focus our comparative studies, we conducted a preliminary nonrepresentative survey among behavioral researchers who use online platforms for conducting research. We distributed our survey through the Society for Judgment and Decision-Making distribution list and through our personal Facebook and Twitter accounts. We received 129 responses, mostly (80%) from researchers in academia. About 40% indicated their discipline as psychology, another 16% as business or management, 15% as judgment and decision-making (or behavioral economics), 5% as political science, and 5% as cognitive sciences (19% did not choose a discipline). All participants indicated some experience with conducting research online, with the median experience being 5 years. We gave respondents descriptions (with examples) of 11 aspects of data quality and asked them to rate how important they personally regard each of them when they choose where to conduct their research online, from 1 (not at all important) to 5 (critically important). The full list of data quality aspects and examples is given in the [Appendix](#). We found that the top three aspects that respondents rated as most important were comprehension, attention, and honesty ($M = 4.41, 4.34, 4.17$, $SD = 0.78, 0.77, 0.89$, respectively), followed by reliability ($M = 3.91$, $SD = 0.9$) and replicability ($M = 3.51$, $SD = 1.09$). The other aspects were rated as important (responses of 4 or 5 on the scale) by less than 40% of respondents. We decided to focus on these top four aspects because they were feasible to implement within a standard survey and would not require splitting our sample into groups of experimental conditions (as is typically required to examine replicability).

Attention refers to the extent to which participants actually read questions before answering them. Typically, this has been measured using attention-check questions (ACQs, also known as instructional manipulation checks, or IMCs; see Oppenheimer et al., 2009) in which the respondent is asked a seemingly benign question (e.g., “Which sport do you like the most?”), but the preface to the question includes instructions to answer in a specific manner (e.g., choose “other” and then type “check”). Another form of ACQs is to embed a nonsensical item within a long scale or questionnaire, to which only one response in the scale's option can be justified (e.g., “When watching TV, I had a

fatal heart attack”; here, attentive participants must answer “Never”; see Paolacci et al., 2010). While attention could also be construed as “latency” (time spent on a question), which is an input measure, we focus our examination on output measures of the performance on ACQs.

Comprehension refers to the degree to which participants seem to understand a task’s instructions (Berinsky et al., 2014; Rand et al., 2012). Although this may seem related to attention, we consider this aspect as testing one level above merely reading instructions. It refers not only to whether participants read the full instructions, but also captures whether they are able to convey instructions clearly back to researchers. For example, this may include whether participants can correctly summarize instructions for an experimental task that asks them to make some judgment, perform some task, or consider different alternatives for a certain goal. Previous research has indeed shown that excluding participants who fail on attention is not the same as excluding those who miscomprehend instructions (Berinsky et al., 2014).

Honesty refers to the extent to which participants provide truthful responses (e.g., to demographic questions) or provide accurate responses when asked to self-report their performance. Extant research on unethical behavior has repeatedly shown that, when given the opportunity, many people will choose to over-report their performance on experimental tasks in order to earn a higher pay (Gerlach et al., 2019). While such dishonesty is prevalent, it can be sometimes detrimental to behavioral research. For example, it has been shown that MTurk participants will sometimes lie about their demographic characteristics in order to claim false eligibility for a study (Chandler & Paolacci, 2017).

Lastly, *reliability* relates to the extent to which participants’ responses (mainly to scale items) are internally consistent and can be reliably used to measure people’s traits, individual tendencies, or preferences. Reliability is considered a prerequisite to the validity of measures and thus can be highly important for behavioral research that aims at understanding people’s attitudes, behavioral intentions, or individual differences in responding to behavioral interventions.

We focused on these four aspects of data quality to inform the design of the following studies that compared the data quality across multiple platforms and panels. We fully recognize that outcomes of interest in data quality may vary and evolve over time, as well as between disciplines. That is why we have conducted the preliminary survey to focus on the above key aspects of data quality that seem most central for current behavioral research.

The current research

Informed by the findings of our preliminary survey, the audiences we first considered included Prolific, MTurk,

CloudResearch (powered by TurkPrime; we considered both their MTurk Toolkit service and their PrimePanels service), Dynata (formerly known as ResearchNow), Qualtrics Panels, YouGov, Lucid, and SurveyMonkey Audience. Our selection criteria included (i) the ability to run the study independently through our software (Qualtrics) and (ii) the ability to grant bonuses to participants (required to test dishonesty). These criteria excluded YouGov, Lucid, SurveyMonkey Audience, and CloudResearch’s PrimePanels. This left us with a final list of five platforms: Prolific, MTurk, CloudResearch’s MTurk Toolkit, Dynata, and Qualtrics Panels. We distinguish between the first three platforms, which are *self-service platforms* that provide researchers complete control over the sampling and administration of their study, from the latter two, which act as *middleman services* of online panels that handle sampling and administration of the study on behalf of the researchers.

The three self-service platforms offer different options for prescreening respondents. On MTurk one can choose Masters only, or limit to workers with high approval ratings (e.g., more than 95% of previous submissions being approved). CloudResearch and Prolific, in addition to approval ratings filters, offer various additional prescreeners (e.g., to block suspicious or duplicate IPs, or use a pre-approved sample of workers). We focused our examination on US respondents, but in order to not limit our sampling further, in Study 1 we did not apply any data quality prescreening filters on any of the sites. To that end, we did not add any requirements on MTurk, we added no prescreening filters on Prolific, and we selected the “All MTurk Participants” option on CloudResearch. Not applying filters was also imperative, so we could include the panels (Dynata and Qualtrics) in Study 1, in which such options are not given at the researcher’s discretion. However, this made the samples of MTurk and CloudResearch conceptually similar, except for the fact that CloudResearch samples from the MTurk pool using “micro-batches” (that allow a faster sampling with lower commission). We henceforth denote the MTurk sample obtained using the CloudResearch interface as “MTurk(CR).” In Study 2, we extended our examination to include prescreening filters and focused only on platforms that offer such filters to researchers and further examined differences in data quality when such prescreening filters are used. Thus, our first study compares two samples of MTurk (one through Amazon’s and the other through CloudResearch’s interface) to samples from Prolific, Qualtrics Panels, and Dynata. Study 2 compares MTurk, Prolific, and CloudResearch.

We pre-registered both studies design and procedure, and these forms, along with all materials and data, are available at <https://osf.io/342dp/>.

Study 1

Participants We requested a sample of 500 participants from each platform, who reported residing in the United States¹, in September 2020. Participants were paid 1.5 GBP on Prolific, 1.5 USD on MTurk and CR, and unknown amounts on Qualtrics and Dynata.² We created no additional demographic quotas other than restricting to 18 years or older. We received a total of 2857 responses, but 347 did not finish the study completely. Table 1 shows the total sample size obtained from each site, the number of complete responses, and key demographic information. Dropout rates differed significantly between the sites, $\chi^2(4) = 206.05$, $p < .001$: They were highest on Dynata and Qualtrics and much lower on MTurk, Prolific, and CR (see Table 1). Note that Qualtrics added 25 respondents on their own volition (with no extra charge).

Among those who completed the study ($N = 2508$), we discovered 79 responses that had the same participant ID that was used in a previous submission (35 in the MTurk sample, 43 in the CR sample, and one in the Prolific sample). This occurred because the study had to appear twice on MTurk: once through our native MTurk account and once through CR. We excluded the responses that were recorded later and kept only the first recorded submission in the final sample. The final sample thus included 2429 participants.

As shown in Table 1, MTurk and CR had fewer women than the other platforms, $\chi^2(4) = 107.14$, $p < .001$. Age differences were also significant, $F(4, 2412) = 357.2$, $p < .001$, especially due to the higher ages of Qualtrics participants (half of the sample were 64 years old or above). Additional demographics are reported in the Appendix. Almost none of MTurk participants used a mobile device to complete the study, whereas the majority of Dynata participants used a mobile device.

Procedure Participants were invited to complete a survey on individual differences in personal attitudes, opinions, and behaviors. All participants began the survey by answering demographic questions, followed by the data quality measures described below. Participants finished the survey by answering questions related to their usage of the online platform including how often they use the site, for what purposes, how much they earn in an average week, their percent of approved submissions (responses that participants submit and are approved by the researcher), and how often (if at all)

Table 1 Description of samples between sites in Study 1

Site	Completed*	Dropout rate (%)	% Women	Mean age (<i>SD</i>)	% Mobile
MTurk	465	11.5	33.5	36.5 (11)	0.4
MTurk(CR)	451	0	33.3	36.9 (11)	1.1
Prolific	499	2.5	52.7	31.5 (11)	19
Qualtrics	525	20.6	47.2	60.4 (15)	25.5
Dynata	489	21.3	60.1	38.6 (16)	59.1

*Without duplicates

they use other sites. These questions were not included for Dynata participants, at the request of Dynata representatives. Overall, the study's average duration was 13.21 minutes ($SD = 14.84$). Regarding the data collection rate, it took us 3:30 hours to complete the sample on MTurk, 2:40 hours on CR, and 2:28 hours on Prolific. Dynata and Qualtrics started with a "soft launch" of 105 and 28 responses, respectively, and then (after requesting our approval) it took Dynata 16:07 to complete the rest of the sample ($N = 384$) and Qualtrics 30:25 to complete the rest of the sample ($N = 499$). Ignoring the soft launch parts, these times indicate that the average response rate per hour was 209.8 on Prolific, 186 on MTurk(CR), 142.6 on MTurk, 23.4 on Dynata, and 16.3 on Qualtrics.

Measures We included distinct measures of each data quality parameter: attention, comprehension, honesty, and reliability. *Attention* was measured with two ACQs embedded within the survey, one more explicit and one more covert (consistent with best practice guidelines; e.g., Keith et al., 2017). The first, more explicit ACQ was presented at the beginning of the survey: participants were presented with a paragraph of instructions followed by two 7-point-scale questions, and were asked to answer "2" on the first question, add 3 to that number, and use that value for the second question (any different response for either question indicated a failure to pass this ACQ). The second, more covert ACQ was a bogus item included within the Need for Cognition (NFC; Cacioppo et al., 1984) scale—"I currently don't pay attention to the questions I'm being asked in the survey." Responses other than "Strongly disagree" were considered a failure to pass this ACQ.

Comprehension was assessed by asking participants to summarize, in their own words, the instructions to two tasks. In the first, participants were told to prepare for a test of their perceptual abilities, in which an image would be presented showing several different faces. They were instructed to count the number of persons they see in the image and report this number within a 20-second period. However, immediately below this was another paragraph, telling them to ignore the instructions above and report seeing zero persons in the image, regardless of the correct answer. Participants were asked

¹ It is important to note that some research has highlighted that even with the use of US-only filter on MTurk, as much as 5–10% of the sample may be recruited from outside the United States (Feitosa et al., 2015)

² The rates were chosen based on the average hourly rate on the platform, which is somewhat higher on Prolific than MTurk and CR. Overall study costs (without bonuses) showed that MTurk and CR were the cheapest options (1050 USD and 1070 USD); Prolific was more expensive at ~1253 USD (975 GBP); Dynata and Qualtrics were more expensive at ~1628 USD (1372.5 EUR) and 2300 USD, respectively.

to summarize these instructions in their own words, and we used their open-ended responses to assess their comprehension of the instructions. Afterwards, an image was displayed showing several faces, and participants were prompted to report the number of faces they see in the image. The second item measuring comprehension was participants' summaries of the instructions to the matrix task that measures dishonesty (see below).

Participants' responses to these two items were coded by two independent reviewers to identify whether participants demonstrated reading and understanding the instructions. Raters were instructed to code any response that suggested a minimum level of understanding as indicating comprehension and only flag responses as incorrect if they were undoubtedly illegible (e.g., "everyoen is poosib;ekl when the right one is near to us", "(", too short (e.g., "nothing", "good," "nice"), irrelevant (e.g., "Paraphrasing means formulating someone else's ideas in your own words...") or a clear copy of the instructions. Raters coded responses blindly, without knowing which site it came from. We found a 94% agreement rate in the first question and 90% in the second. We resolved disagreements using a conservative approach under which only responses that were flagged by both raters were coded as incorrect answers.

Honesty was measured using an online version of Mazar et al.'s (2008) matrix problem task, which appeared at the end of the survey. In each problem, participants were presented with a table containing 12 numbers between 0 and 10 with two decimal digits (e.g., 3.48, 5.12). Their task was to find the two numbers in each table that added up to exactly 10, within 20 seconds. Participants were told that they would earn a small bonus reward for each problem they reported as solved (0.1 GBP on Prolific, 0.1 USD on all other platforms). Five problems were presented, but the fifth problem was actually unsolvable. We focused on this problem's reports as a measure of dishonesty³.

Reliability was measured using the NFC scale (Cacioppo et al., 1984), which measures the extent to which respondents like to engage in and enjoy thinking, and the Domain-Specific Risk-Taking Scale (DOSPRT; Blais & Weber, 2006), which evaluates self-reported risk-taking behavior. The NFC contains 18 items and was measured on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). The negatively worded items were recoded before analyses. The DOSPERT contains 30 items, scored on a 7-point scale from 1 (extremely unlikely) to 7 (extremely likely). We chose these

scales because they have been found as highly reliable and were validated across extensive studies previously.

Results

We first report differences between sites on each data quality aspect separately, and then aggregate these findings to a composite score of data quality between sites.

Attention We found statistically significant differences between sites in both ACQs, $\chi^2(4) = 548.48, 203.56, p < .001$. Figure 1 shows that Prolific had the highest rate of passing ACQs with 68.7% passing both ACQs, compared to 46.6% and 45.5% on MTurk(CR) and MTurk. Qualtrics and Dynata had the lowest average rates of passing ACQs with only 22.5% and 22.1% (respectively) passing both ACQs. As Fig. 1 shows, there was a larger difference between the sites on the first, more explicit ACQ, on which MTurk and MTurk(CR) did as well as Prolific, but Qualtrics and Dynata did very poorly. In contrast, the differences between the sites were smaller on the second, more covert ACQ.

Comprehension Raters found 1034 (43%) incorrect answers to the first (faces) question and 434 (17.9%) in the second (matrix) question. As verification of the coding of the first (faces) question, we also observed that 92% of those who incorrectly summarized the instructions also failed to follow them (did not choose "zero persons" as instructed; also, 86% whose summaries were correct indeed acted as instructed). There were statistically significant differences between the sites, as well as between the questions, as can be seen in Fig. 2. On the first question, 82% of Prolific participants answered it correctly, compared to only about half of the participants on the other sites, $\chi^2(4) = 152.4, p < .001$. On the second question, 98% of Prolific participants answered it correctly, compared to 69% from MTurk or MTurk(CR) and 87% on Qualtrics and 85% on Dynata, $\chi^2(4) = 215.5, p < .001$. Overall, 81% of Prolific participants answered both questions

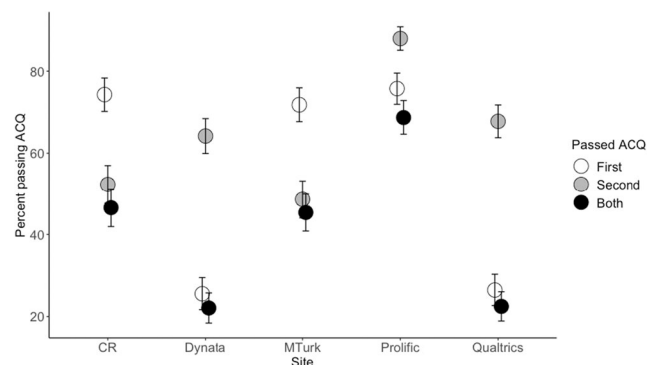


Fig. 1 Attention rates between sites (error bars show 95% confidence intervals)

³ We report one deviation from our pre-registration: another question that was aimed to measure dishonesty asked participants whether they want to participate in a study that recruited participants outside their reported age range, but we encountered a technical error, and the age range in the question was not displayed correctly to most participants. The question was successfully used in Study 2.

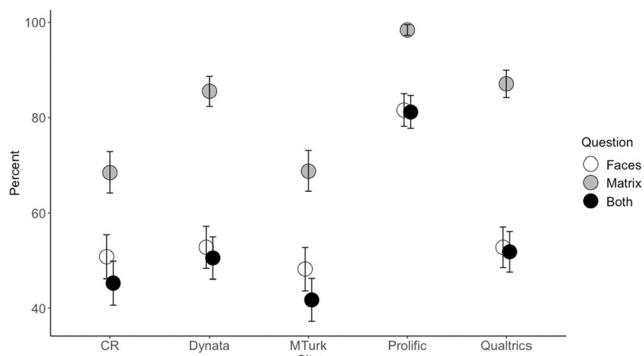


Fig. 2 Comprehension rates between sites (error bars show 95% confidence intervals)

correctly, compared to 45% and 42% on MTurk(CR) and MTurk, and 52% and 51% on Qualtrics and Dynata, respectively, $\chi^2(4) = 193.91, p < .001$

We next examined the correlation between passing ACQs to answering correctly the comprehension questions. As can be seen in Fig. 3, participants who passed both ACQs achieved much better scores on the comprehension questions compared to those who failed them, across all sites. A logistic regression with site and attention (passing both ACQs) on comprehension (correct answers to both questions) showed significant effect for attention, $OR = 18.05, 95\% CI [11.41, 29.26], p < .001$. The differences between the sites were also significant, as MTurk(CR)'s success rates were 1.83 times larger than MTurk's, Prolific's 9.18 times larger than MTurk, and Qualtrics and Dynata 4.43 and 4.34 times larger than MTurk, respectively, $ps < .001$. We also found significant interactions between attention and site: $OR = 0.44, 0.35, 0.42, 0.34, ps < .05$, for MTurk(CR), Prolific, Qualtrics, and Dynata (all compared to MTurk), respectively. This interaction suggests that lower attention decreased comprehension on all sites, but mostly for MTurk and MTurk(CR), and less so on Prolific, as can be seen in Fig. 3.

Honesty Across the five matrix problems, MTurk and MTurk(CR) respondents claimed the highest proportion of

problems solved (54.5%, 56.1%, respectively), followed by Dynata (42.1%), and Prolific and Qualtrics (36.9%, 32.6%, respectively). This translates to an average difference of \$0.23 in bonus claimed per participant between the highest (MTurk[CR]) and lowest (Qualtrics) groups. To directly assess honest behavior, we examined the rate of participants who *did not* claim to have solved the last matrix problem, which was in fact unsolvable. We found statistically significant differences between the sites, $\chi^2(4) = 153.44, p < .001$. About 84% of Prolific users were honest on this problem, followed by Qualtrics (78%), Dynata (69%), and MTurk/MTurk(CR) (55%). We then contrasted honesty levels between participants who passed both ACQs vs. those who failed (one or both) ACQs. As Fig. 4 shows, participants who passed ACQs were more honest than those who failed ACQs (79% vs. 62% on average across sites). A logistic regression analysis on cheating rate with attention (passing both ACQs) and site as predictors showed that, compared to MTurk, Prolific and Qualtrics' honesty rates were much higher ($OR = 3.59, 95\% CI [2.34, 5.61], OR = 3.80, 95\% CI [2.73, 5.33]$, respectively). Dynata's honesty rate was also higher than MTurk's ($OR = 2.33, 95\% CI [1.69, 3.24]$), but MTurk(CR)'s cheating rate was not significantly different from MTurk's ($OR = 0.92, 95\% CI [0.65, 1.32], p = 0.66$). Higher attention (passing ACQs) increased honesty considerably ($OR = 2.66, 95\% CI [1.82, 3.91]$). The interactions between the sites and attention were not significant ($p > .62$), suggesting that the relative effect of attention was similar between all sites, which is indeed consistent with the rates in Fig. 4.

Reliability Table 2 details the internal reliability (Cronbach's alpha) for all scales, between sites, and between participants who passed or failed the ACQs. For the NFC scale, we found high reliability coefficients (0.79–0.94) except among MTurk and MTurk(CR) participants who failed on attention. In these two groups, reliability was zero or negative, indicating no correlation between responses to the items of the scale. For

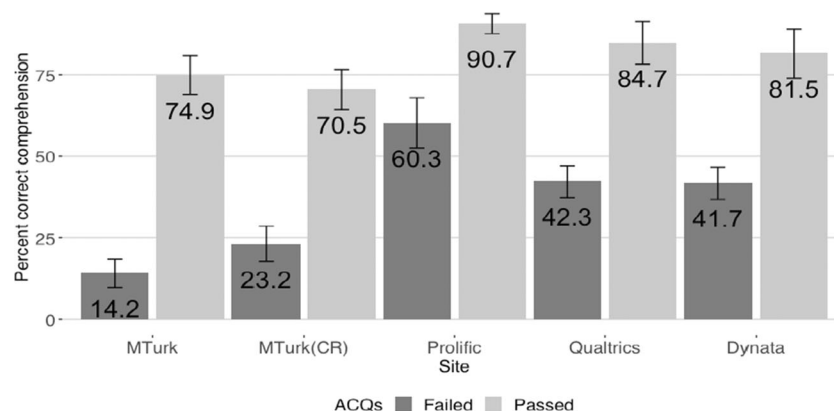


Fig. 3 Comprehension rates between sites and levels of attention

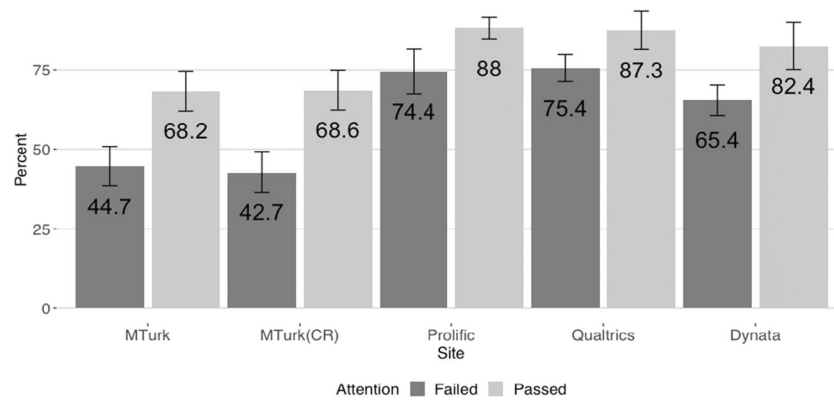


Fig. 4 Honesty rates between sites and levels of attention (error bars show 95% confidence intervals)

the DOSPERT scale, we found some differences between sites and attention levels. First, for the MTurk samples, we found high reliability across all five sub-scales, with small differences between participants who failed or passed ACQs (except for the Social sub-scale, in which participants who passed ACQs produced lower reliability). For Prolific, Dynata, and Qualtrics, we found that reliability was actually higher among participants who *failed* ACQs than among those who passed them. For Dynata and Qualtrics, this difference might be attributed to the fact that the size of the groups of participants who passed ACQs in those samples was smaller, compared to those who failed ACQs, but that is not the case with the Prolific sample.

Overall data quality scores We computed, per participant, an overall composite score of data quality based on the individually measured aspects of attention (passing ACQs), comprehension, and dishonesty (reliability cannot be used, as it is an aggregated rather than individual-level score). The score gave participants a value between 0 and 5, showing whether they passed one or both ACQs, answered correctly one or two

comprehension questions, and did not claim to have solved the unsolvable problem. The correlations between the five measures ranged between 0.16 and 0.43, all $p < .01$, but the overall composite score should not be considered as measuring the same construct. Rather, it is used here as a multifactorial measure that attests to the overall general level of data quality. Figure 5 shows the average scores between the sites, and between participants who used a desktop or mobile device (except MTurk/MTurk(CR), where too few used mobile devices).

As can be seen in Fig. 5, participants on Prolific had the highest overall score among all sites. The differences between the sites were significant, $F(4, 2420) = 81.84$, $p < .001$. Pairwise comparisons, with Bonferroni correction, showed that the difference between Prolific and all other sites was significant, $p < .01$, whereas the differences between all other sites were not, $p > .33$. The effect size of the difference between Prolific and the other sites was Cohen's $d = 3.70$, indicating a very large difference in overall data quality. The effect of the device (examined only among the three sites that had enough mobile users) was also significant, $F(1, 1507) =$

Table 2 Internal reliability (Cronbach's α) between the sites and attention

Site	Attention	N	NFC	DOSPERT				
				Ethical	Financial	H/S	Rec.	Social
MTurk	Passed	211	0.94	0.90	0.87	0.84	0.89	0.69
	Failed	254	-0.12	0.88	0.88	0.88	0.87	0.79
MTurk(CR)	Passed	210	0.93	0.90	0.85	0.82	0.89	0.67
	Failed	241	0.07	0.87	0.86	0.89	0.87	0.83
Prolific	Passed	343	0.92	0.68	0.73	0.63	0.80	0.65
	Failed	156	0.86	0.91	0.91	0.86	0.87	0.72
Qualtrics	Passed	118	0.93	0.75	0.73	0.51	0.66	0.68
	Failed	407	0.85	0.92	0.89	0.88	0.91	0.76
Dynata	Passed	108	0.90	0.54	0.55	0.65	0.77	0.56
	Failed	381	0.79	0.89	0.88	0.83	0.87	0.75

H/S health/safety, Rec. recreational

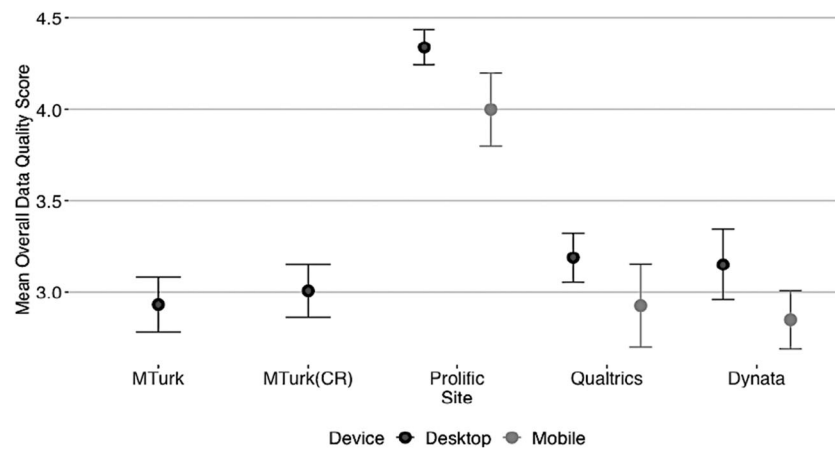


Fig. 5 Overall data quality score between sites and devices

16.85, $p < .001$, without a significant interaction, $p = 0.92$. This effect showed that in all those sites, desktop participants exhibited higher data quality than mobile users. The effect size of the device was Cohen's $d = 0.43$, indicating a medium-sized difference in overall data quality.

Usage patterns Many participants on MTurk(CR) and MTurk (42% and 41%, respectively) said they use the site as their main source of income, in contrast to only 8% on Prolific or 7% on Qualtrics. We found significant differences in the frequency of usage between the sites. As can be seen in Fig. 6, Qualtrics participants reported the lowest frequency of usage, as about 29% of them indicated they use the site less than 1 hour a month, and nearly two thirds use it less than 2 hours per

week. In contrast, MTurk and MTurk(CR) showed the highest frequency of usage, as many of them reported using the site at least 8 hours a week or more. Most Prolific participants reported a medium level of usage mostly ranging between 1 and 8 hours a week, as shown in Fig. 6. The differences between frequency of usage distributions were significant between sites, $\chi^2(21) = 823.84$, $p < .001$.

Participants' self-reported approval rates (the percent of their submissions that were approved) were similar on average between the sites: $M = 90.5$, 89.6 , 90.3 , $SD = 20.1$, 22.6 , 20.1 , for MTurk, MTurk(CR), and Prolific, respectively (we did not ask Qualtrics and Dynata participants this question). Medians were at 98% for all three sites, suggesting the majority of participants from all sites had very high approval rates.

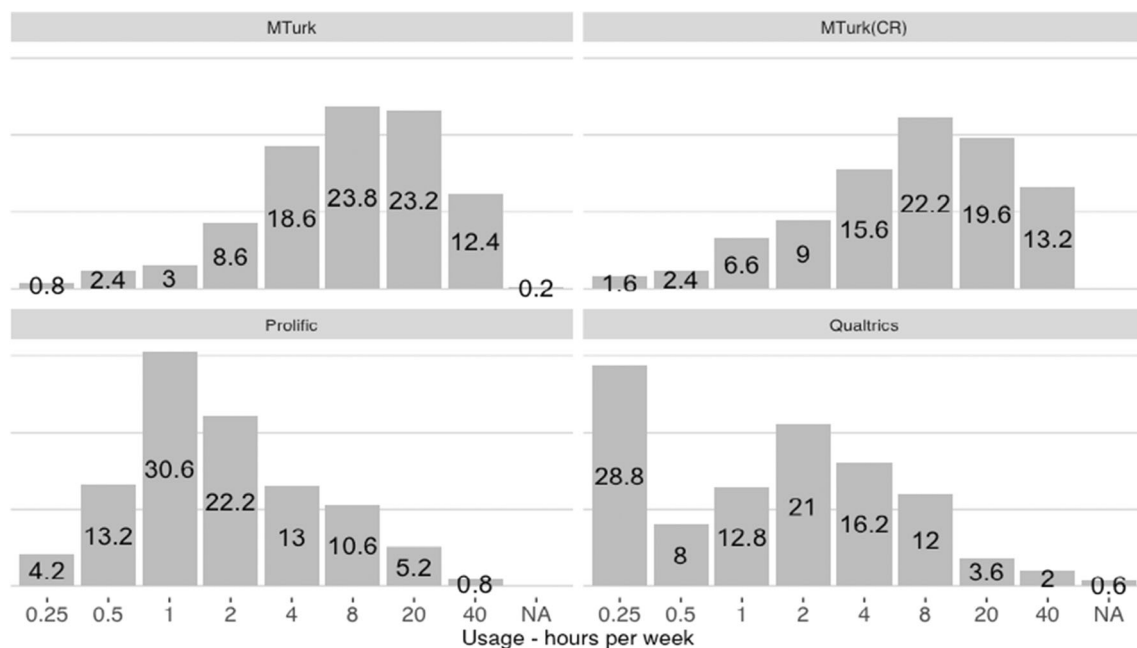


Fig. 6 Distribution of frequency of usage (percent within site) between sites. Note: 0.25 and 0.5 hours/week means up to 1 or 2 hours per month, respectively

Table 3 Percent of responses from each site indicating using other sites (multiple-choice question: total percentages can exceed 100% across each row)

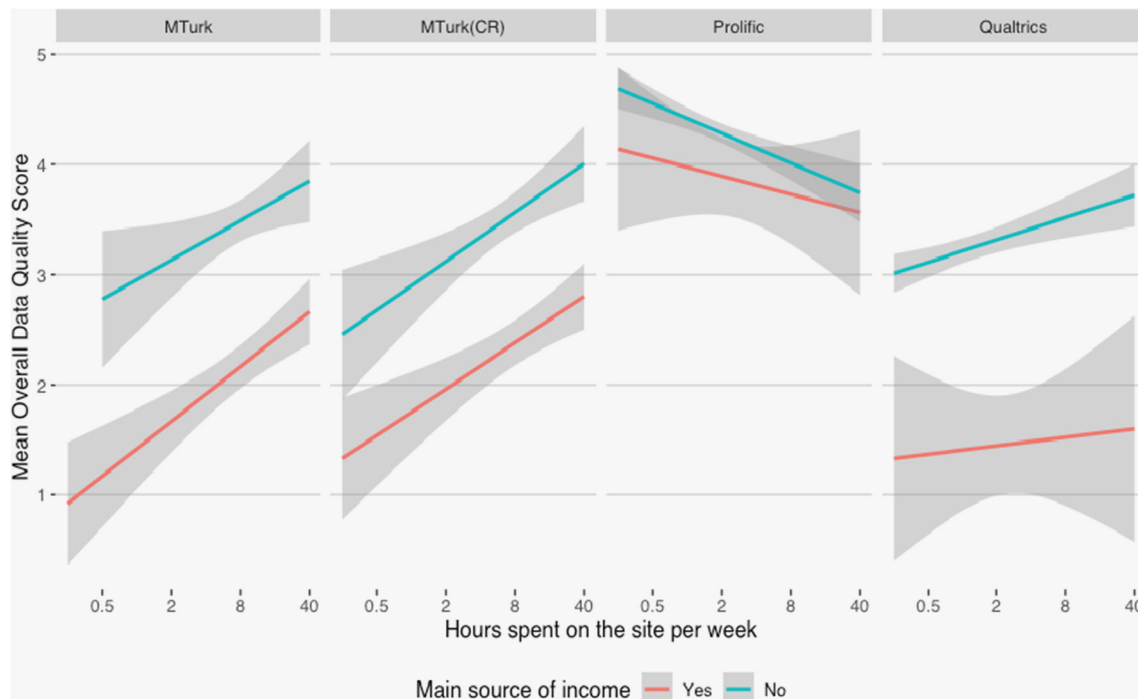
Responses from:	Qualtrics (%)	Prolific (%)	MTurk (%)	Dynata (%)	No other platform (exclusive users) (%)
MTurk	34	43.2	–	45.2	34.4
MTurk(CR)	43.1	45.5	–	32.4	36
Prolific	14.4	–	21.2	4.2	69.6
Qualtrics	–	4.2	4.8	21.3	76.6

Regarding usage of other platforms, Table 3 shows the percentages of participants in each site indicating they have used the other sites “several times” or “many times.” As detailed, more than 40% of MTurk/MTurk(CR) participants report to also use Prolific, and at smaller but similar rates use Qualtrics or Dynata. In fact, less than 40% of participants from MTurk or MTurk(CR) said they use no other platform other than MTurk. Prolific and Qualtrics participants, in contrast, mostly reported using only their platform, although 21% of Prolific participants also reported using MTurk and 21% of Qualtrics participants also reported using Dynata. These differences in the distributions were statistically significant, $\chi^2(12) = 1602.13, p < .001$.

Usage and data quality To examine whether usage patterns predict data quality, we focused on the frequency of using the site (hours per week) and whether the site was participants’ main source of income, because these questions were measured on four out of the five sites we examined (except Dynata) and showed significant differences between the sites

(compared to approval ratings, which were unavailable for Qualtrics and also not statistically different between the other sites). ANOVA showed significant effects for site, $F(3, 1872) = 120.73, p < .001$, hours spent on the site, $F(7, 1872) = 7.81, p < .001$, and whether the site was the main source of income, $F(1, 1872) = 246.93, p < .001$. We also found significant interactions between site and hours, $F(21, 1872) = 5.04$, site and main source of income, $F(3, 1872) = 8.63$, and hours and main source of income, $F(7, 1872) = 3.55$, all $p < .001$. The three-way interaction was not significant, $F(20, 1872) = 1.35, p = 0.14$. To further explore this, we ran a linear regression model with site as a factorial fixed effect and hours as a continuous predictor. Figure 7 shows the predicted regression lines for data quality as a function of hours spent on the site, and whether the site is the main source of income between the sites.

Between the sites, we found that, compared to MTurk, Prolific had a significant positive effect on data quality, $b = 0.79$ ($SE = 0.08$), $p < .001$, while MTurk(CR) and Qualtrics were not significantly different from MTurk, $b = 0.05, 0.01$

**Fig. 7** Predicted data quality by hours spent on site and whether site is main source of income, between sites (with standard errors)

($SE = 0.08$), $p = 0.52$, 0.89 , respectively. We also found a significant effect for the hours, $b = 0.18$ ($SE = 0.06$), $p < .001$. This suggests a positive relationship between hours spent on the site and data quality ($\beta = 0.26$). However, there was a significant interaction between the site being Prolific and hours spent on the site, $b = -0.42$ ($SE = 0.08$), $p < .001$, suggesting that for Prolific the relationship was actually the inverse. Indeed, that can be observed in Fig. 7. The interaction terms for MTurk(CR) and Qualtrics with hours spent on the site were not significant, $b = 0.03$, -0.14 ($SE = 0.06$, 0.08), $p = 0.72$, 0.08 , respectively, suggesting, as shown in Fig. 7, that these sites had a similar positive relation as found for MTurk.

We also found a significant main effect for the site being the main source of income, $b = -1.75$ ($SE = 0.5$), $p < .001$, showing that data quality was lower among those who indicated the site as their main source of income, by a magnitude of -0.40 SDs. This can be also observed in Fig. 7. Note, however, that for Prolific and Qualtrics, the percent of participants indicating the site as the main source of income was less than 10%. Still, although the interaction between the site being the main source of income and the site being Prolific was significant, $b = 0.29$ ($SE = 0.07$), $p < .001$, data quality was still higher on average in Prolific for participants who said it was their main source of income versus not, $t(45.94) = 2.61$, $p = 0.01$.

Screen time and straight-lining Lastly, we examined differences in the time participants spent on some of the questions for which we collected timing measures (the Faces task and the NFC and DOSPERT scales), as well as the tendency for “straight-lining” (providing identical responses to items on the NFC and DOSPERT scales). The results of these analyses are reported in the [Appendix](#) (for both this study and Study 2).

Discussion

Among all the considered alternatives, Prolific provided data with the highest quality on almost all measures. Prolific participants devoted more attention to the questions of the study, comprehended instructions better, answered questionnaire items carefully, and behaved more honestly when given the opportunity to cheat to increase their gains. While this serves an endorsement of the data quality from Prolific’s participants in this sample, it also shows the tests we designed were reasonable for people to answer correctly, which make the failure rates among the other sites concerning. There is a general lack of attention among the respondents for both of the two MTurk-based samples and the two panels.

Regarding reliability, an extremely low reliability was found for the NFC scale among MTurk participants who failed the ACQs. One explanation for this finding may relate

to the fact that the NFC contains negatively worded items, while the DOSPERT does not. This may suggest that inattentive MTurk respondents completed the NFC without reading all the questions but with trying to provide a consistent profile of responses (e.g., all high or all low). The DOSPERT, which does not contain negatively worded items, showed different findings in which for MTurk participants there was no difference between those who failed or passed ACQs (both groups provided high reliability), but there were differences between those groups among the other samples. Thus, in Study 2 we re-examined this issue using only the NFC, as it contains both positively and negatively worded items.

The differences in attention, comprehension, and dishonesty between the sites were considerable and could not be explained by differences in reputation (self-reported approval ratings) between the resulting samples, because the vast majority of participants had high approval ratings and we found no significant differences between the sites on this measure. This is especially interesting because of the fact that we did not restrict our samples, in any of the sites, to include only highly reputable participants. As previously mentioned, we intentionally did not use any prescreening filters on any of the sites or services sampled in this study. We are aware, however, that the recommendation on some sites, as well as from previous literature (e.g., Peer et al., 2014), is to apply filters to sample only those with high reputation, and it is indeed possible that such settings could yield different results. We test that in Study 2.

Differences in usage patterns, especially in relation to hours spent on the site per week and whether participants use the site as their main source of income, did predict data quality in a nontrivial manner. First, we found that in the MTurk-based samples, the more participants used the site, the higher their data quality was. If one considers higher frequency of usage to suggest less naivety to research, this finding is interesting in light of previous research findings that have found that non-naivety can hamper data quality, and especially replicability (Chandler et al., 2015). While we did not test replicability directly, it can be assumed that replicability of known effects requires that participants pay attention, comprehend, and provide honest answers to a study’s questions. However, another explanation could be that MTurk participants with more experience are more skilled at identifying ACQs or more familiar with dishonesty checks. Furthermore, it is also somewhat puzzling how this relationship was actually inverse among the Prolific sample, where more active participants displayed lower data quality. We retest these findings in Study 2.

Furthermore, participants (mostly from the MTurk-based samples) who indicated they use the site as their main source of income exhibited considerably lower data quality. One possible explanation could be that such participants care more about completing as many studies as possible, to increase their gains, and thus pay less attention to each individual study.

Coupled with the findings about frequency of usage, our findings point to an especially careless group of participants (in the MTurk samples), who state that they use the site as their main source of income but spend relatively few hours on the site. These participants might be especially motivated to maximize their gains out of each study they complete and might do so by paying less attention and cheating more. We also revisit this finding in Study 2, in which we add another measure of specific dishonest behavior of claiming false eligibility to a high-paying study (Chandler & Paolacci, 2017).

Payment rates on the different platforms and panels could explain some of the differences in the results (e.g., Buhrmester et al., 2016). As previously mentioned, we paid MTurk and CR participants 1.5 USD and Prolific participants 1.5 GBP in order to make payments fair relative to the common hourly rate on the different platforms. On Qualtrics and Dynata, in contrast, we could not know the actual payment rates given to participants. Thus, our current study was not properly designed to examine any potential effects of payment rates on data quality, and future research should aim to consider this aspect more directly. In Study 2, we adjust payment rates to be more equivalent.

The middleman services we examined, Qualtrics and Dynata, drawn from online panels exhibited very similar data quality, even though the demographic composition of the samples was quite different. Regardless, it appears that these two sites provided data quality that was comparable and sometimes a little better than the quality of the MTurk samples, even though it was still much lower than Prolific's. However, sampling from these services took much longer (about 5–6 times more), their costs were higher, and more overhead was required in coordinating and running the study. Thus, they appear to be an inferior alternative in our choice set, as they are the most expensive (both in terms of money and time), with no marginal benefit in quality over the others. Because of that, and because of our desire to add prescreening filters to the samples, we did not include panels in Study 2.

Study 2

Generally, Study 2 used the same design and procedure as Study 1, with two major changes: (i) we focused only on MTurk, CloudResearch's MTurk Toolkit (henceforth, CR) and Prolific, and (ii) we added prescreening filters to all samples to maximize their potential data quality and better simulate what many researchers are likely to apply when doing research on these platforms. We again tested the four key aspects of data quality that we found to be most important for behavioral researchers in our pre-study, namely attention, comprehension, reliability, and dishonesty. However, in light of Study 1's findings, we limited our examination of reliability

to one scale and expanded our examination of dishonesty to include more than one kind of measure.

Participants We recruited 500 participants from each platform (MTurk, CR, and Prolific), who reported residing in the United States, in March 2021. Participants were paid 1.5 USD on CR and MTurk and 1.1 GBP on Prolific plus a bonus of up to 0.5 USD/GBP. We applied data quality prescreening filters on all sites by restricting the study to participants with at least 95% approval rating and at least 100 previous submissions; on CR we also used the site setting to “block low data quality workers.” We excluded participants on Prolific who completed the previous study on Prolific and participants on CR who completed the previous study on CR or MTurk. However, because our study had to be posted twice on MTurk (once through our MTurk account and once through our CR account), 39 participants completed the study twice, and we removed their later submission (although they were still paid for their submissions). The final sample thus included 1461 participants who completed the study. Table 4 presents the samples. Additional demographics can be found in the [Appendix](#).

Design and procedure The study was identical to Study 1 except for the following changes: (1) in the first ACQ, the values participants were asked to enter in response were changed to 6 and 3; (2) in the matrix task, we included two (instead of one) unsolvable problems; (3) we omitted the DOSPERT scale, keeping only the NFC scale, as it included negatively worded items; (4) we added an “imposter” question to the dishonesty measures. The “imposter” question, which came at the very end of the study, asked participants whether they would like to be invited to a study in the future. We told participants that the future study investigates a specific subpopulation of people, and thus it offers higher pay than usual (“up to 15 USD per hour”). Participants were told that the study is open to participants who are male/female (according to the gender of the participant, which they indicated at the beginning of the study) and are at a given age range, which was programmed to be from 5 to 9 years older than the age participants reported in the beginning of the study (e.g., for a person who said they were 30 years old, the age range for

Table 4 Description of samples between sites in Study 2

Site	Completed*	Dropout rate (%)	% Women	Mean age (SD)	% Mobile
CR	489	9.90	52.8	41.1 (12.5)	3.90
MTurk	472	15.40	36	36.9 (10.5)	1.30
Prolific	500	5.50	52.6	35.7 (13)	19

*Without duplicates

recruitment was 35–44). Participants could choose to say they (1) fit the criteria and wanted to take part, (2) fit the criteria but did not want to take part, (3) did not fit the criteria, or (4) other. Responses of 1 were coded as dishonestly claiming false eligibility. Lastly, we also included a question after the usage questions asking participants whether they had participated in this study or a similar study in the past. The study took participants on average 9.8 minutes ($SD = 5.2$). It took 4:34 hours to complete data collection on CR, 4:57 hours on MTurk, and 5:31 hours on Prolific.

Results

Attention and comprehension Figure 8 presents the percentage of participants from each site who passed each ACQ and each comprehension questions. In the ACQs, both CR and Prolific showed higher passing rates than MTurk. Testing on the percentage who passed both ACQs, the overall differences were significant, $\chi^2(2) = 124.51, p < .001$, as was the difference between CR and Prolific, $\chi^2(1) = 17.15, p < .001$, indicating that CR participants performed better on attention in this study than Prolific, who performed better than MTurk. In the comprehension questions, raters showed a 98% agreement and, as in Study 1, we coded responses as failing the questions only if both raters gave it a score of zero. As can be seen in Table 4, both CR and Prolific passed both questions significantly more often than MTurk, $\chi^2(2) = 160.01, p < .001$. The difference between CR and Prolific was not significant, $\chi^2(1) = 1.62, p = 0.2$.

We examined how attention predicted comprehension using a logistic regression with passing both ACQs and site as the predictors for passing both comprehension questions. We found a significant negative effect for the site being MTurk (vs. CR as the reference point), $OR = 0.11, 95\% CI$

[0.04, 0.26], $p < .001$, but no significant effect for the site being Prolific, $p = 0.62$. Passing ACQs had a significant positive effect on comprehension, $OR = 2.94, 95\% CI [1.22, 6.59], p < .001$. Similarly, the interaction between attention and the site being MTurk was also statistically significant, $OR = 4.38, 95\% CI [1.72, 11.82], p < .001$. The interaction of Prolific with attention was not significant, $p = 0.77$. These results suggest that while higher attention led to higher comprehension in all sites, it was mostly effectual for the MTurk sample and not so for Prolific or CR. In other words, the marginal gain from adding ACQs on Prolific and CR appears to be small.

Dishonesty MTurk participants claimed the highest percent of problems solved (49%, $SD = 30.9$), followed by CR (38%, $SD = 29.9$) and Prolific (32%, $SD = 24.8$). In the two unsolvable matrix questions, 71% on Prolific were completely honest (did not claim to solve either of them) compared to 69% on CR and 46% on MTurk, $\chi^2(4) = 88.46, p < .001$. The specific difference between Prolific and CR was also significant, $\chi^2(2) = 7.39, p = 0.02$, suggesting that Prolific participants were more honest than participants on CR and MTurk. In the “imposter” question, 60% of MTurk participants claimed false eligibility, compared to 55% on CR and 48% on Prolific, $\chi^2(2) = 12.69, p < .01$. The specific difference between CR and Prolific was also significant, $\chi^2(1) = 4.06, p = 0.04$, showing that Prolific participants were more honest than CR and MTurk. Figure 9 shows that attention (passing both ACQs) also predicted honest behavior, as the percent of participants who neither cheated on either of the two unsolvable problems nor claimed false eligibility for a future study was higher if they passed ACQs versus failed them (35% vs. 14%), $\chi^2(1) = 42.94, p < .001$. The difference was most pronounced among MTurk users, where only 3% of those who failed ACQs were fully honest.

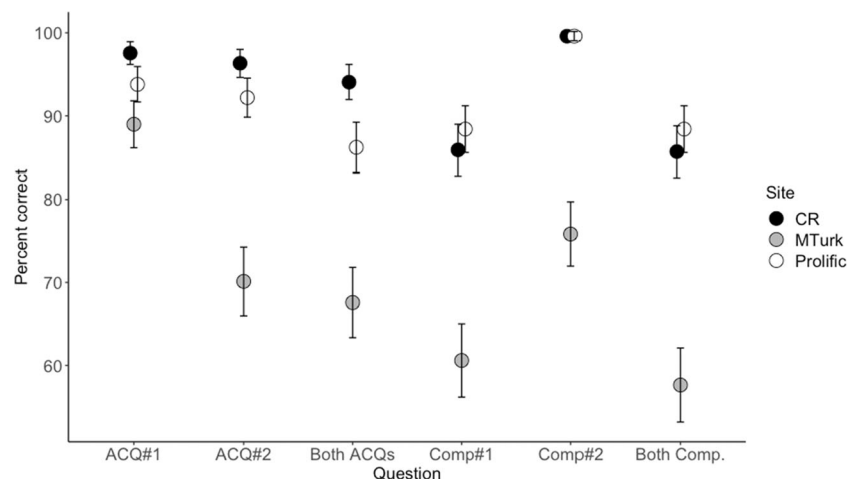


Fig. 8 Attention and comprehension in Study 2 between sites

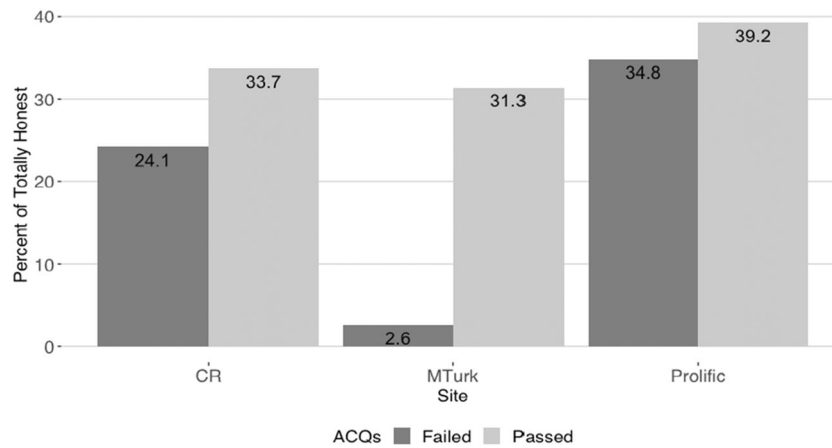


Fig. 9 Percent of participants who did not cheat at all between sites

Reliability Cronbach's alpha for the NFC was high among MTurk, CR, and Prolific participants who passed both ACQs (0.95, 0.96, 0.94, respectively). For those who failed the ACQs, reliability was still high on CR and Prolific (0.91, 0.89, respectively), but less than adequate among MTurk participants who failed the ACQs (0.33). This finding is consistent with the low reliability found among MTurk participants in Study 1 for the NFC scale.

Overall data quality score Similar to Study 1, we computed an overall data quality score per participant by aggregating whether they passed one or both ACQs, answered correctly one or both comprehension questions, did not claim to solve one or both unsolvable matrix problems, and did not claim false eligibility for a future study. This composite score ranged from 0 to 7 and showed an average of 5.41 ($SD = 1$) and a median of 6. We found statistically significant differences between the sites on this measure, $F(2, 1458) = 129.4$, $p < .001$, which showed higher scores for Prolific and CR ($M = 5.87$, 5.78 , $SD = 1.0$, 1.1 , respectively) compared to MTurk ($M = 4.55$, $SD = 1.9$). Post hoc tests with Bonferroni correction showed that the differences between Prolific and MTurk and between CR and MTurk were significant, $p < .001$, but the difference between CR and Prolific was not, $p = 0.91$.

Usage and data quality Thirty percent of MTurk participants said they use the site as their main source of income, whereas that rate was 12% on CR and 4% on Prolific, $\chi^2(2) = 130.4$, p

$< .001$. Table 5 shows self-reported overlap between the sites: as many as 23% of Prolific users said they also use MTurk, while more than 30% of MTurk workers said they also use Prolific.

Similar to Study 1, we found that MTurk and CR participants use the site more hours per week than Prolific participants (see Fig. 10): while the median response on MTurk and CR was spending 8 hours per week, it was only 2 hours per week on Prolific, $\chi^2(14) = 441.05$, $p < .001$. Figure 11 shows predicted data quality scores by frequency of usage and whether the site is the main source of income for the participant. A regression analysis with site, hours, and main source as predictors of overall data quality showed a negative main effect for the sample being from MTurk, $b = -2.15$ ($SE = 0.24$), $p < .001$, but no significant main effects for hours or main source, $p > 0.6$. We did find a significant interaction between the site being MTurk and hours, $b = 0.53$ ($SE = 0.19$), $p < .01$, suggesting that the effect of hours was evident only on the MTurk sample. The three-way interaction of MTurk, main source, and hours was also significant, $b = 1.71$ ($SE = 0.24$), $p < .001$, suggesting the following significant trend shown in Fig. 8: for participants on MTurk who use it as their main source of income, their data quality increases the more hours they spend on the site; but for those who do not use the site as the main source of income, numbers of hours spent does not increase data quality. On CR and Prolific, it appears that hours spent on the site and/or using

Table 5 Percent of responses from each site indicating using other sites (multiple-choice question: total percentages can exceed 100% across each row)

Site	MTurk (%)	Prolific (%)	SurveyMonkey (%)	CrowdFlower (%)	Other (%)	No other platform (exclusive users) (%)
CR	–	25.2	11.9	1.8	3.3	66.7
MTurk	–	32.2	27.3	17.8	14.6	51.7
Prolific	22.8	–	7.0	0.4	4.8	69.8

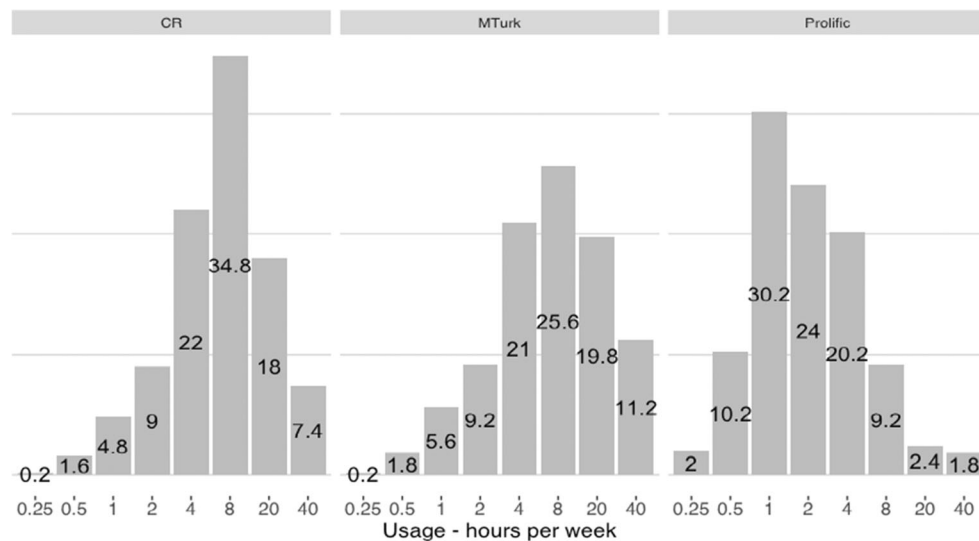


Fig. 10 Frequency of usage (hours spent per week) between the sites

the site as the main source of income does not predict any differences in overall data quality.

Reputation and data quality We found significant differences in self-reported approval ratings between the sites, as CR participants reported higher approval ratings ($M = 99.14$, $SD = 1.6$) than MTurk or Prolific ($M = 96.91, 96.16$, $SD = 8.0, 5.0$, respectively), $F(2, 1355) = 38.74$, $p < .001$. A regression analysis of overall data quality with site and approval ratings as predictors showed a significant positive effect for approval rating, $b = 0.11$, $SE = 0.04$, $p < .01$, but no significant interactions with the sample, $p > 0.05$.

Discussion

The goal of this study was to replicate and extend the examination of differences in data quality between the sites in Study 1 to a situation where a researcher samples from each platform using common prescreening filters of data quality (e.g., approval ratings). Using these filters, our findings suggest that the native MTurk sample (with the approval ratings filter) offers data quality that is considerably inferior to that obtained through either CR (which samples from the same pool) or Prolific (which does not). MTurk participants exhibited lower attention and comprehension, and cheated considerably more, than participants on CR or Prolific. Regarding reliability on

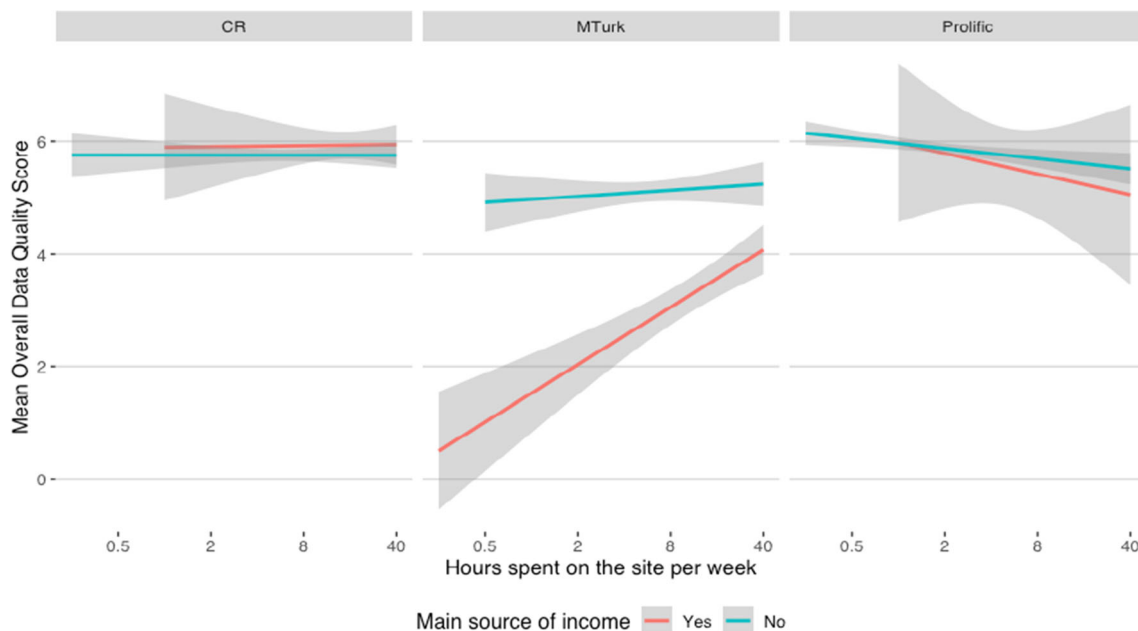


Fig. 11 Predicted overall data quality between sites as a function of hours spent on the site and whether the site is the main source of income (with standard errors)

validated scales, MTurk participants who failed on attention also were unable to produce adequate reliability, which was high among all other groups. Thus, the two first conclusions of Study 2 are that (i) MTurk is vastly inferior to both CR and Prolific, and (ii) the effect of the attention-check questions was restricted to MTurk.

Between Prolific and CR, there seemed to be an advantage for CR participants in relation to passing ACQs, but an advantage to Prolific participants in all the questions that tested honesty. In other words, while CR participants were more attentive than Prolific participants, they also took more opportunities to act dishonestly. However, despite these differences being statistically significant, both platforms showed high rates of attention (especially compared to MTurk), and the difference in cheating was also not very large. In addition, the platforms showed comparable results on the other aspects of comprehension and reliability, and had very similar overall data quality scores. Thus, it can be concluded that when the data quality prescreening filters are on, data quality from both Prolific and CR is similarly high.

Comparing the overall results between Study 1 (without filters) and Study 2 (with filters) confirms the claim (also raised in Litman et al., 2021) that using CR without their prescreening filters is actually similar to sampling directly from MTurk, and CR's higher data quality relies on researchers using their "block low data quality" feature. This means that it is specifically this unique feature implemented by CR that drives data quality, and not approval rating (the latter was used on both CR and MTurk). This again suggests that previous recommendations to rely only on reputation (approval rating) as a sufficient condition for data quality (Peer et al., 2014) are challenged, at best. In that former study, restricting only to high-reputation workers resulted in not needing to use ACQs. In contrast, the results of the current study suggest that even when restricting to high-reputation workers, data quality can be less than sufficient, requiring researchers to exclude participants who fail ACQs.

For Prolific, there appears to be no noticeable difference between data quality with (Study 2) versus without (Study 1) filters. This might be because approval ratings are used less on Prolific, because researchers do not reject enough bad submissions, or because the majority of participants already have high approval ratings. Additionally, Prolific participants provide a host of demographic data upon signing up to the platform, and that could also be part of the reason they provide high data quality. In any case, researchers using Prolific would probably benefit from not applying filters that are too strict, because they can reduce the sampling frame and might also reduce the sample's representativeness. However, further research is needed to evaluate whether such a trade-off actually exists and to what degree.

Regarding usage patterns and their effect on data quality, we again found a considerable difference between the MTurk-

based samples and those of Prolific, whose participants seem to use the site much less, and very few use it as their main source of income, compared to a much higher usage pattern among both CR and MTurk participants. Moreover, for the MTurk sample, these two factors contributed considerably to the observed (lower) data quality there: workers who used the site less but claimed it was their main source of income exhibited much lower data quality than their counterparts. This confirms the concern raised in Study 1 that these respondents constitute an especially precarious group that could greatly hamper the overall data quality of a study and should be avoided. The most straightforward way to do so would be to use CR or Prolific and avoid using Amazon's native MTurk site.

General discussion

Across both studies, our findings suggest that although considerable improvement has been made in the market of platforms and panels for online research, data quality remains a concern that researchers must deal with before deciding where to conduct their online research, and also should take into account when configuring the sampling frame. Apparently, online panels and platforms cannot be a priori considered as equivalent in terms of the data quality they may provide. Specifically, it appears that some panels (Qualtrics and Dynata) cannot provide adequate data quality that would be of interest to behavioral researchers: many of their participants fail attention-check questions, do not seem to comprehend instructions, and cheat considerably when given the opportunity. Furthermore, these panels seem to be much more expensive than MTurk, CR, or Prolific. Thus, they seem to be inferior options on both quality and price. However, further research is needed to examine whether and to what degree such panels provide more representative samples, and whether that could justify the higher cost and lower quality they seem to provide.

The low data quality of the native MTurk sample is, in our opinion, even more disturbing. MTurk is probably the most commonly used platform for behavioral researchers, despite its relatively high commission (40% currently) and limited user interface (compared to CR or Prolific). Even after applying data quality filters (approval ratings, number of submissions, and location), MTurk's data quality was inferior to CR and Prolific's on almost all the key aspects that we measured: they failed ACQs more, more frequently erred on comprehension questions, and cheated more often (than Prolific), and reliability was adequate only among those who passed ACQs. This could lead to a suggestion to use MTurk with good ACQs, but that would mean that a large proportion of the sample would have to be excluded from analyses. Thus, we do not recommend that researchers use the MTurk native

site. However, if MTurk is the only option for data collection, there are a number of recommendations for enhancing response validity that should be considered, such as via the use of standardized reporting checklists (for comprehensive reviews of such recommendations see Cheung et al., 2017; DeSimone et al., 2015; Keith et al., 2017). These recommendations are likely to be useful when using other platforms as well and should be encouraged, as our results show that even in the best-performing samples (e.g., Prolific and CR with filters) there can still be some low-quality responses.

Judging both from the responses to our preliminary survey and our communal knowledge and experience, it appears that more researchers use MTurk than Prolific or CR. Lower costs or higher familiarity of MTurk can explain this difference, but additional surveys are needed to better understand why different researchers use different platforms and panels, and how much, if at all, they consider data quality in their decision-making. Our preliminary survey was not conducted on a representative sample and could have included other data quality aspects or other formats (e.g., open-ended explorative questions), and thus its findings are highly limited. We therefore argue it is important to further examine researchers' preferences and considerations to deepen the examination of data quality aspects in the future.

Additional research should also tackle the trade-offs between data quality and naivety more directly. In order to obtain a sample with strong data quality from MTurk, CloudResearch samples respondents who spend a lot of time on the system, presumably conditioned towards popular behavioral experiments and surveys. Building from the baseline of students in a lab, behavioral researchers did not mark naivety as a top attribute, but it can be a critical attribute for many fields of research, and it is possible that behavioral research will expect more of it in the future. Additional research should also consider other ways to measure data quality using, for example, different types of ACQs and other measures for dishonesty.

On that note, we hope that this research encourages a more frequent review of both the aspects of data quality important in different domains, and the data quality exhibited by the leading options for gathering respondents. We expect that data quality concerns could be different across research domains, change over time as research evolves, and depend on other factors, such as the tools different platforms offer researchers to help improve data quality. The various options for gathering responses also evolve at a rapid pace with new modes and companies each year. Thus, a consistent framework, like the one we present in this paper, which makes online research easily replicable and comparable, is necessary to ensure quality and trust in this type of research.

Appendix

List of data quality aspects rated in the preliminary survey (in alphabetical order):

- **Attention** - whether and to what extent participants seem to devote enough attention to answering the questions (e.g., pass attention-check questions)
- **Attrition**- rates of participants not returning to follow-up stages of a longitudinal study.
- **Comprehension**- whether and to what degree participants seem to understand the question's instructions (e.g., summarize them correctly, follow their instructions).
- **Drop-out rate** - amount of participants not completing the study.
- **Honesty**- whether and to what extent participants provide truthful responses (e.g., to demographic questions), or provide accurate responses when asked to self-report their performance.
- **Naivety**- whether and to what extent participants are not overly familiar with common research materials and tasks.
- **Reliability**- to what extent participants provide internally consistent responses (e.g., test-retest measures, internal consistency on validated scales).
- **Replicability**- whether and to what extent the samples from this platform have been found to provide effect sizes that are consistent (in direction and significance) with findings from previous original studies.
- **Representativeness**- how representative the sample is of desirable target populations (e.g., on demographic and geographic variables).
- **Response speed**- the time by which a desired sample size can be obtained.
- **Thoroughness**- the depth, detail, and elaboration that participants typically provide when responding to open-ended questions.

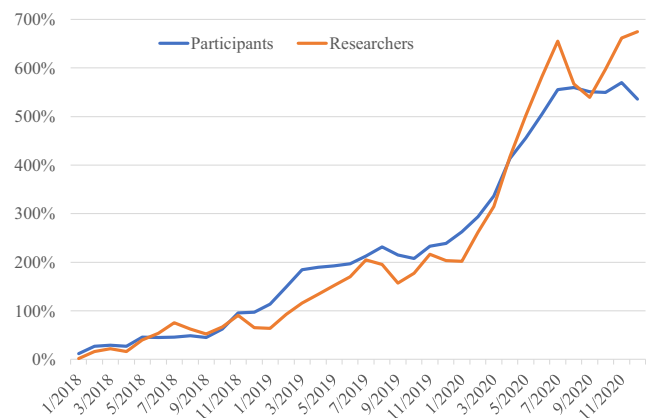


Fig. 12 Percent increase in number of participants and researchers on Prolific, compared to end of 2017, by month (data courtesy of Prolific).

Table 6 Percentage of participants indicating as each ethnicity across platforms (study 1)

Site	White	African American	Asian	Latin/Hispanic	Other	Prefer not to answer
CR	62.3%	27.9%	3.3%	5.5%	0.4%	0.4%
MTurk	67.1%	25.8%	2.6%	3.7%	0.9%	0%
Prolific	54.7%	10%	24.2%	8%	2.4%	0.6%
Dynata	75.3%	11.9%	5.3%	4.9%	1.6%	1%
Qualtrics	86.1%	2.1%	6.3%	2.5%	0.8%	2.3%

Table 7 Percentage of participants indicating yearly income within each band across platforms (study 1)

Site	Under \$10k	\$10–25k	\$25–50k	\$50–75k	\$75–100k	\$100–150k	Over \$150k	Prefer not to answer
CR	3.8%	12.2%	33%	33%	12.6%	3.3%	1.3%	0.7%
MTurk	9%	16.4%	23.1%	16%	14.9%	10.4%	6.1%	4.1%
Prolific	3.9%	10.8%	35.5%	30.5%	14.6%	3.7%	0.4%	0.6%
Dynata	18%	14.4%	21.8%	15.2%	10.4%	8.4%	4.4%	7.2%
Qualtrics	4.2%	7.2%	17.9%	15.8%	14.5%	20.4%	13.1%	6.9%

Table 8 Percentage of participants indicating highest earned degree across platforms (study 1)

Site	Some High School	High School	Assoc Degree	Bachelors	Some Grad School	Masters	Other Prof. Degree	PhD	Other	Prefer not to say
CR	0%	10%	7.5%	50.8%	3.3%	25.1%	0.9%	2%	0%	0.4%
MTurk	0.2%	10.1%	6.2%	52.5%	3.4%	25.2%	0.9%	1.1%	0.2%	0.2%
Prolific	1.2%	25.7%	10%	32.1%	6.2%	17.8%	2.4%	3%	1.2%	0.4%
Dynata	3.1%	30.1%	17.4%	22.5%	6.5%	11%	1.8%	4.3%	1.8%	1.4%
Qualtrics	1.5%	20.4%	13.5%	26.3%	6.7%	21.5%	2.5%	4.4%	1.9%	1.3%

Table 9 Percentage of participants indicating as each ethnicity across platforms (study 2)

Site	White	African American	Asian	Latin/Hispanic	Other	Prefer not to answer
CR	83%	6.7%	5.9%	2.9%	1.4%	0%
MTurk	72.7%	18.4%	3.8%	2.8%	1.5%	0.8%
Prolific	73.4%	2.6%	14%	7.2%	2.4%	0.4%

Table 10 Percentage of participants indicating yearly income within each band across platforms (study 2)

Site	Under \$10k	\$10–25k	\$25–50k	\$50–75k	\$75–100k	\$100–150k	Over \$150k	Prefer not to answer
CR	6.5%	14.9%	33.3%	19.2%	14.1%	7.2%	2.7%	2%
MTurk	5.9%	11.9%	28.6%	35.8%	9.5%	4.7%	2.1%	1.5%
Prolific	15%	15.4%	27%	16.2%	12.8%	6.8%	3.6%	3.2%

Table 11 Percentage of participants indicating highest earned degree across platforms (study 2)

Site	Some High School	High School	Assoc Degree	Bachelors	Some Grad School	Masters	Other Prof. Degree	PhD	Other	Prefer not to say
CR	0.8%	22.7%	15.1%	39.9%	4.7%	14.3%	1%	1.2%	0.2%	0%
MTurk	0%	15.7%	10.4%	47.9%	5.3%	19.3%	0.4%	0.6%	0.2%	0.2%
Prolific	0.8%	25.6%	12.8%	35.2%	5%	14.8%	2.6%	1.6%	1.6%	0%

Table 12 Time spent (in seconds) on selected questions between sites in studies 1 and 2.

Site	Faces task (Md)	NFC scale (Md)	DOSPERS scale (Md)
Study 1			
MTurk	71.54	43.63	65.62
MTurk(CR)	72.37	41.67	62.77
Prolific	80.83	106.22	128.11
Qualtrics	60.38	100.18	120.71
Dynata	64.27	105.23	122.96
<i>Significance test</i>	$\chi^2(4) = 76.65, p < .001$	$\chi^2(4) = 463.56, p < .001$	$\chi^2(4) = 488.94, p < .001$
Study 2			
CR	73.63	76.86	
MTurk	71.04	59.06	
Prolific	73.65	96.92	
<i>Significance test</i>	$\chi^2(2) = 1.95, p = 0.37$	$\chi^2(2) = 151.46, p < .001$	

Table 13 Straight-lining (maximum identical rating*) on the scales in studies 1 and 2 between sites and levels of attention (passing vs. failing attention-check questions)

	NFC Study 1		DOSPERS Study 1		NFC Study 2	
	Failed ACQs	Passed ACQs	Failed ACQs	Passed ACQs	Failed ACQs	Passed ACQs
MTurk	51.6	45.9	46.1	46.4	49.5	45.2
CR	51.6	45.7	46.9	44.7	46.0	45.6
Prolific	48.9	43.3	45.5	43.0	50.7	44.7
Qualtrics	52.0	45.7	59.8	59.7	-	-
Dynata	54.0	43.7	50.4	54.8	-	-

*The maximum identical rating method identifies which value is most commonly used by the respondent and then calculates the percentage of items that value was selected by the respondent. The measure ranges from 0 (least straightlining) to 100 (most straightlining) for a given respondent (Kim et al., 2019, page 218).

Declarations

Disclosure We wish to acknowledge the fact that the research project was done in collaboration with members from Prolific (ED is co-founder and CEO of Prolific; DR serves on the Prolific advisory board; ZE and AG are employed by Prolific), and the study was funded by Prolific. However, we have taken several precautions to mitigate any potential conflict of interest. First, the study was pre-registered on the Open Science Framework, and all materials and data are available at <https://osf.io/342dp>. This enables outside researchers to corroborate our analyses and findings, as well as easily conduct replication studies or extensions of this study to other platforms or aspects of data quality. Second, the studies on all platforms were run by the first author (EP) who is not affiliated with Prolific in any way and did not receive any financial or other compensation. EP was also responsible for planning the research design, analyzing the data, drafting, revising and submitting the final manuscript. Other authors' contributions were as follows: DR took part in the research design and revising the surveys, pre-registering the study, checking the data analyses and revising the manuscript; ZE assisted with the research design, conducting the preliminary survey, and writing the manuscript; AG ran statistical analyses and contributed to writing the manuscript; ED assisted with the research design and provided feedback on the manuscript.

References

- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3), 800.
- Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3), 739–753.
- Blais, A. R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1), 33–47.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1), 3–5.
- Buhrmester, M. D., Talafar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13(2), 149–154.
- 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. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5), 500–508.
- 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. A. (2015). Using nonnaïve participants can reduce effect sizes. *Psychological Science*, 26(7), 1131–1139.
- Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51(5), 2022–2038.
- Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, 32(4), 347–361.
- DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, 36(2), 171–181.
- Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, 75, 47–52.
- Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, 145(1), 1.
- Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, 44(1), 196–210.
- 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.
- Göriz, A. S., Borchert, K., & Hirth, M. (2021). Using attention testing to select crowdsourced workers and research participants. *Social Science Computer Review*, 39(1), 84–104.
- Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in non-probability online panels. Online panel research: A data quality perspective, 1, 219–237.
- Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. *NYU Working Paper No. CEDER-10-01*, Available at SSRN: <https://ssrn.com/abstract=1585030>. Accessed 1 Sep 2021.
- Keith, M. G., Tay, L., & Harms, P. D. (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, 8, 1359.
- Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–webmixed-mode surveys. *Social Science Computer Review*, 37(2), 214–233.
- Litman, L., Moss, A., Rosenzweig, C., & Robinson, J. (2021). Reply to MTurk, Prolific or panels? Choosing the right audience for online research (January 28, 2021). Available at SSRN: <https://ssrn.com/abstract=3775075>. Accessed 1 Sep 2021.
- MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. J. (2018). The accuracy of measurements with probability and nonprobability survey samples: replication and extension. *Public Opinion Quarterly*, 82(4), 707–744.
- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6), 633–644.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872.
- 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.
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163.
- Porter, C. O., Outlaw, R., Gale, J. P., & Cho, T. S. (2019). The use of online panel data in management research: A review and recommendations. *Journal of Management*, 45(1), 319–344.
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427–430.
- Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, 8(2), 206.

- Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77, 184–197.
- Thompson, A. J., & Pickett, J. T. (2019). Are relational inferences from crowdsourced and opt-in samples generalizable? Comparing criminal justice attitudes in the GSS and five online samples. *Journal of Quantitative Criminology*, 36, 907–932.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980–991.
- Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpsen, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709–747.

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