Reply to MTurk, Prolific or panels? Choosing the right audience for online research

Leib Litman [1,2], Aaron J. Moss [2], Cheskie Rosenzweig [2,3], and Jonathan Robinson [2,4]

[1] Department of Psychology, Lander College, Flushing, NY
[2] CloudResearch, Queens, NY
[3] Department of Clinical Psychology, Columbia University, New York, NY
[4] Department of Computer Science, Lander College, Flushing, NY

Disclosure: This article is a reply to a previous article published on SSRN by Peer et al, (2021) on which four of the five authors are members of Prolific. As the members of Prolific did in their paper, we wish to clearly state that all authors of this commentary are members of CloudResearch (LL and JR are co-founders and Chief Research Officer and Chief Technology Officer, respectively; AM and CR are both employed by CloudResearch). The research reported here was funded by CloudResearch. This study used the exact same stimuli that were used by Peer at al. to replicate their CloudResearch findings. The one difference between the studies is that we left the CloudResearch data quality settings on. The stimuli we used and the data we collected is open and available on OSF. We strongly encourage any interested researchers to replicate the findings reported here. All authors contributed to and approved of the final version of this comment.

Abstract

In a recent paper published on SSRN, Peer et al., (2021) compared data quality across five participant recruitment platforms commonly used for research in the behavioral sciences. After finding evidence to suggest Prolific data is superior to alternatives, the authors, who are themselves primarily members of Prolific, state that using other platforms "appears to reflect a market failure and an inefficient allocation or even misuse of scarce research budgets" (pg., 21). Such an assertion has the potential to change how research funds are allocated in potentially harmful ways. Therefore, we sought to interrogate the claims made by Peer et al. We found surprising methodological decisions, which were undisclosed in their paper, that severely limit the inferences that can be drawn from their data. Most notably, when these researchers gathered data with the CloudResearch MTurk Toolkit, they chose to turn off the recommended data quality filters, including filters that are on by default and were designed to address known data quality issues on MTurk. When we replicated their study using these recommended options, we found CloudResearch data superior to that of Prolific. After presenting our findings, we discuss several theoretical factors that are crucial for evaluating the strengths and weaknesses of different online platforms and encourage researchers to adopt a "fit for purpose" view when evaluating platforms for online data collection.

Reply to MTurk, Prolific or panels? Choosing the right audience for online research

In a recent paper published on SSRN, Peer, Rothschild, Evernden, Gordon, and Damer (2021) compare data quality on multiple online participant recruitment platforms that are commonly used for scientific research in the behavioral sciences. Among the platforms included in their study were Mechanical Turk, CloudResearch, two market research panels, and the platform the authors are affiliated with—Prolific. They found Prolific to be "superior to other options," and even go so far as to state that the use of platforms other than Prolific "appears to reflect a market failure and an inefficient allocation or even misuse of scarce research budgets" (pg., 21). Such an assertion has the potential to change how research funds are allocated in potentially harmful ways. For example, if the reviewers of grants or granting agencies took the statements of Peer et al. at face value they may decide not to fund proposals that intend to use platforms other than Prolific for data collection. Because different online platforms have unique strengths and weaknesses (e.g., access to participant populations not available elsewhere), such a decision would severely restrict the type of data researchers can gather online and possibly impede the efficient accumulation of information about human behavior. For these reasons, we believed it was important to interrogate the claims made by Peer et al., (2021).

Considering the unique characteristics of different online platforms is part and parcel to appropriate research methodology. Surprisingly, when we examined the methods of Peer et al. we found that they did not disclose how they sampled from the five different platforms in their study nor were the reasons for any particular sampling decisions discussed. This is important because failing to disclose such decisions potentially obscures this information from the scrutiny of peer-review.

When we examined the details of the Peer et al sample that was gathered on the CloudResearch site we found that they decided to switch off the data protection tools CloudResearch designed to ensure high quality data. These tools constitute the default settings for data collection on CloudResearch, and, contrary to best practices, these default settings were not used in the Peer et al study. Turning off data quality filters and then considering low quality data as an indictment of a platform is akin to knocking down the walls of a house and then being upset about getting cold and wet. More consequentially, sampling from CloudResearch without the default data quality tools is the same as sampling from Mechanical Turk directly. Thus, rather than comparing the data quality of MTurk and CloudResearch, Peer et al. effectively sampled MTurk twice—something that explains why their numbers for CloudResearch and MTurk were virtually identical.

In this reply, we replicate the Peer et al. study while using the CloudResearch "Approved Group", a feature that CloudResearch recommends for collecting high quality MTurk data when using the CloudResearch platform. Using the exact same stimuli as Peer et al., we find that, when used as intended, CloudResearch data is notably better than what was reported by Peer et al for Prolific and vastly superior to what Peer et al reported for CloudResearch.

Method

Participants

We recruited 500 participants from MTurk using the CloudResearch Toolkit. We conducted a "soft launch" for 200 people to ensure the survey was working properly and that there were no problems with data collection. The next morning, we collected the final 300 responses. Our data are available at https://osf.io/q36ht/ .

Procedural Background

CloudResearch (Litman, Robinson, & Abberbock, 2017) connects to Mechanical Turk through application programming interface (API) integration. If a researcher conducted a study with the CloudResearch Toolkit and removed the data quality tools and other features CloudResearch has built over time (Litman, Rosenzweig, & Moss, 2020) it would be the same as running the study directly on MTurk. This is essentially what Peer et al did. However, in addition to removing CloudResearch's tools, Peer et al., also chose not to apply Mechanical Turk's reputation qualifications (i.e., worker approval rating and past HITs completed). This decision is particularly puzzling given that the lead author of their paper has written a paper advocating the use of these very qualifications as a way to maintain data quality on MTurk (e.g., Peer et al., 2014). None of these methodological details were reported in the Peer et al paper, but were accessible to us on the CloudResearch site.

In contrast to Peer et al., we assessed data quality of the CloudResearch Toolkit using the CloudResearch-Approved group of participants. As a reminder, all this required was that we not change the default settings. The Approved group of participants is a CouldResearch-vetted pool of people on MTurk. The Approved group has more than 50,000 people in the U.S. and mirrors the demographics of the MTurk platform.

Measures

We used the measures from Peer et al., (2021) and exactly replicated their study with one exception. For one item assessing comprehension, the measures posted on Open Science Framework (OSF) by Peer et al. was missing a photo intended to serve as a distractor task. Because of this omission we chose our own photo but kept the wording identical. Other than this

change, the methods were exactly the same. We report our results in the same format as Peer et al.

Attention. The survey included two attention check questions. In the first attention check participants were presented with a paragraph of instructions followed by two Likert-type questions on a 7-point scale. The question text asked people about preferences for employment, but the paragraph of instructions told people to ignore these questions and to "*please answer 'two*' on the first question, *add three to that number and use the result as the answer on the second question*." Therefore, anyone who answered anything other than "two" for the first item and "five" for the second item failed the check.

The second attention check was embedded within the Need for Cognition scale. Specifically, within this 18-item measure one item read, "I currently don't pay attention to the questions I'm being asked in the survey." Responses other than "Strongly disagree" were counted as a failure.

Comprehension. Comprehension was assessed by asking participants to summarize the instructions for two different tasks. On the first task, participants were presented with two short paragraphs of instructional text. The text told participants that in the next task they would have to count the number of people in a photo and quickly report their answer. Further reading, however, told people to ignore the task and report zero for their answer. Participants were asked to summarize the instructions in their own words. Following Peer et al we assessed comprehension by coding the open ended responses and examining participant behavior on the subsequent counting task. Anyone whose summary failed to mention that they should ignore the task and answer zero and anyone who provided an actual answer other than zero were considered to not have adequately read and understood the task.

The second comprehension task was simpler. People were asked to summarize the instructions of a matrix counting task that assessed dishonesty (see below).

Honesty. The honesty measure was a version of Mazar et al.'s (2008) matrix problem task. In each problem, participants are presented with a table that contains twelve numbers each carried to two decimal places (e.g., 3.48, 5.12). All numbers in the table were between 0 and 10. Participants' job was to find the two numbers in each table that add to exactly 10, within 20 seconds. There were five problems in this task and participants were told they would receive a 10 cent bonus for each problem solved. Unbeknownst to participants the final problem was unsolvable. Hence, we examined whether people said they solved the unsolvable task as a measure of honesty.

Reliability. Two well-validated and commonly used measures assessed reliability of participant responses: the Need for Cognition Scale (NFC; Cacioppo, Petty, &Kao, 1984) and the Domain-Specific Risk-Taking Scale (Blais & Weber, 2006). The NFC has 18 items measured on a five point Likert-type scale (1 = *strongly disagree*; 5 = *strongly agree*). The risk-taking scale has 30 items measured on a seven point Likert-type scale (1 = *extremely unlikely*; 7 = *extremely likely*).

Results

Attention

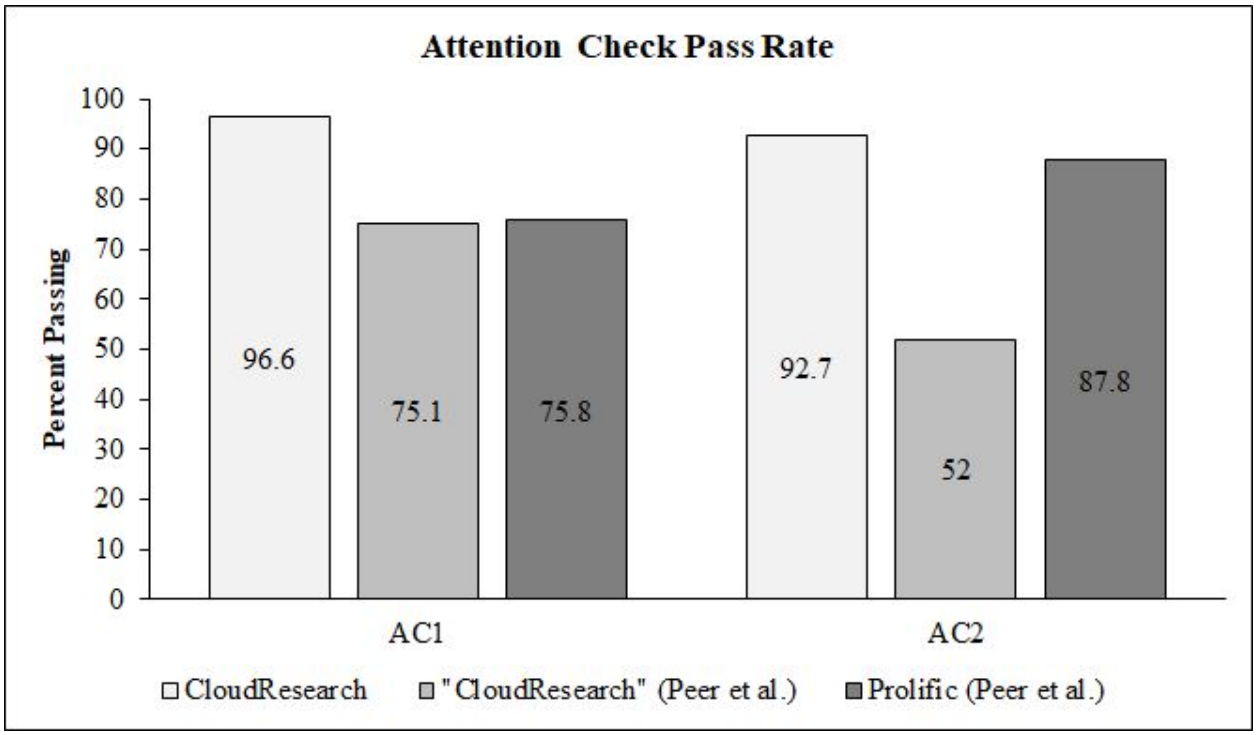
The first attention check within the study had two parts. If participants read the instructions, they should have answered 'two' to the first part and 'five' to the second. Only people who answered both items correctly were deemed to have passed the check. Among

CloudResearch-Approved participants 96.6% of people passed this check (i.e. correctly answered both questions).

For the attention check embedded within the NFI, 92.7% of CloudResearch-Approved participants passed the check.

Comparing the data we gathered to Peer et al. shows that CloudResearch participants in our sample performed better on these items than the Prolific sample and the “CloudResearch" sample gathered by Peer et al. (see Figure 1). (Note: when referring to Peer et al’s findings “CloudResearch” appears in quotations).

Figure 1. Pass rates on attention check items.



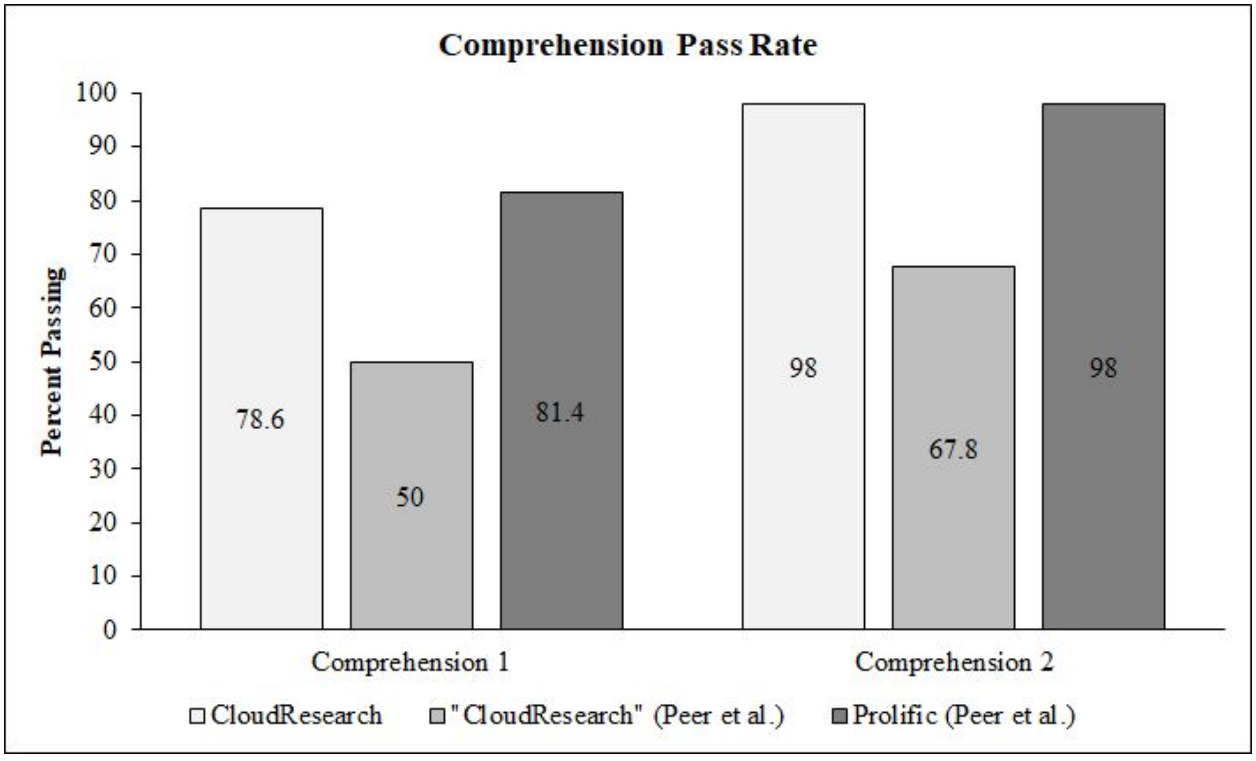
Note: AC = Attention check. AC1 is the percentage of people who passed both items that were part of the attention check. AC2 is the percentage of people who responded ‘strongly disagree’ to the item embedded within the Need for Cognition Scale. Unlike Peer et al., we do not report an average for attention check pass rates.

Comprehension

Rating participants' summary of instructions for the face counting question, we found 78.6% of participants displayed evidence of reading and comprehending the instructions. As validation of our coding judgments, the percentage of people who actually answered 'zero' on the question was almost identical (78.8%). These percentages are substantially higher than the "CloudResearch" numbers reported in Peer et al., and on par with the Prolific numbers (see Figure 2).

The second comprehension question asked people to summarize the instructions for the matrix problem task (Mazar et al., 2008). This task was apparently easier as 98% of CloudResearch participants provided a correct response. In Peer et al., 98% of participants from Prolific also passed this item while the numbers for "CloudResearch" and MTurk were near 70%.

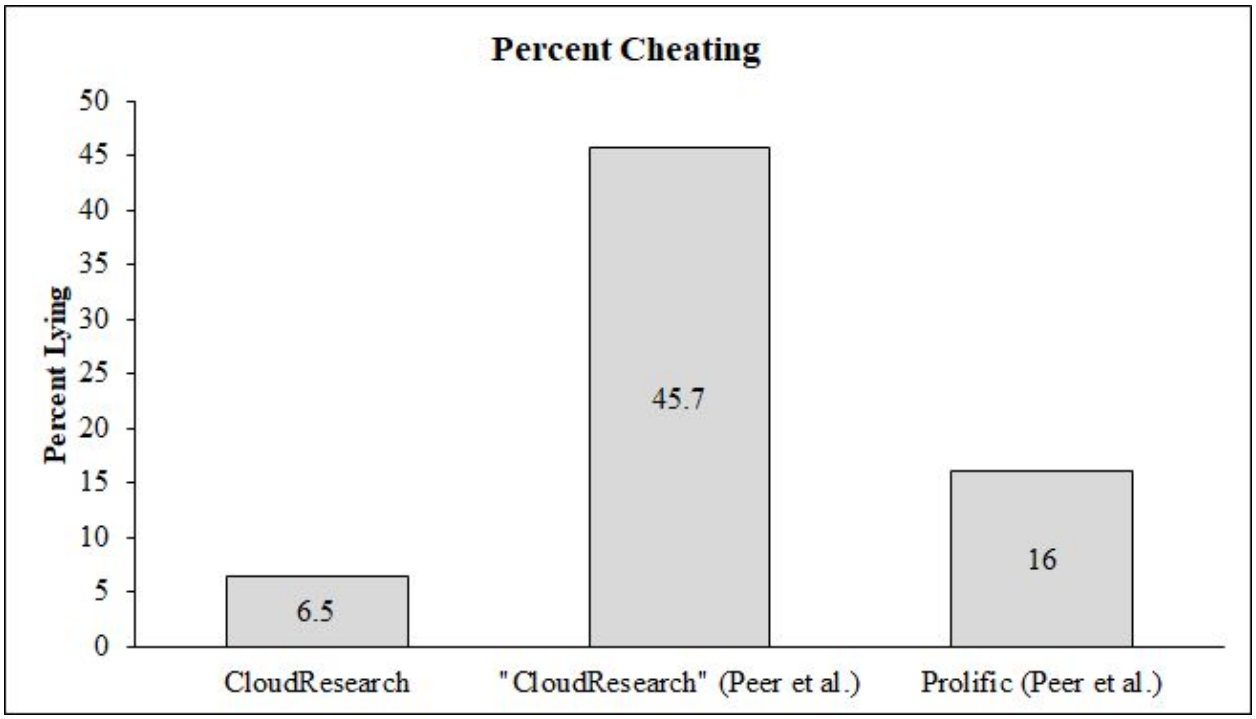
Figure 2. Percent passing the comprehension checks.



Honesty

The key measure of honesty was participants' responses to the final item in the matrix problem. The problem was unsolvable so anyone claiming to have found a solution was likely lying. Only 32 people in the CloudResearch-Approved group said they solved this problem (6.5% of the sample). This number is noticeably lower than any other platform reported in Peer et al., including Prolific. Hence, CloudResearch participants demonstrated the greatest level of honesty (see Figure 3).

Figure 3. The percentage of participants who lied about solving an unsolvable task.



Reliability

Need for Cognition (NFI).

We computed an alpha reliability coefficient for the NFI. The reliability coefficient was .84. Across the five platforms included in Peer et al., NFI coefficients ranged from .82 to .90.

Domain-Specific Risk-Taking (DOSPERT).

The reporting in Peer et al. is not complete enough for us to understand exactly how reliability scores for the Risk-Taking scale were computed. As reported in Blais & Weber, 2006, the Domain-Specific Risk-Taking Scale consists of five subscales. Peer et al report one reliability coefficient for the entire measure, leading us to believe they created one scale with all items included.

We followed the published literature and computed a separate reliability coefficient for each subscale (see Table 1). Each coefficient was in line with past publications and slightly higher than the coefficients reported by Blais and Weber, (2006). When we follow the presumed approach of Peer et al. we obtain a reliability score within the range reported across the five platforms (alpha = .89).

Table 1.

	CloudResearch	Blais & Weber, 2006
Ethical	.75	.67
Financial	.79	.78
Health/Safety	.74	.70
Recreational	.82	.75
Social	.72	.76

Discussion

The data presented above stand in stark contrast to that reported by Peer et al., (2021). In particular, the estimates of data quality when using the default settings of the CloudResearch Toolkit are significantly better in our study than in Peer et al. The CloudResearch Toolkit was developed through years of research during which we designed specific tools to address known and emerging threats on MTurk. CloudResearch has made these tools the default option precisely because researchers have been unable to combat these issues with other methods (e.g., the widespread rejection of suspect submissions). The clear reason for the divergence between Peer et al.'s data and ours is that they chose to remove CloudResearch's default data quality settings. And, as this study shows, when these tools are applied researchers can gather data that is superior in quality to all the platforms assessed in Peer et al., including Prolific.

Importantly, the conclusions of Peer et al. are not only misleading about CloudResearch; they are also misleading about Mechanical Turk. MTurk was designed so that workers and requesters can develop tools and extensions that meet their needs. The CloudResearch Toolkit is one such extension. The CloudResearch Toolkit was built to meet the needs of social science researchers (Litman et al., 2017), including addressing issues of data quality. Because the CloudResearch Toolkit is an extension of Mechanical Turk it is impossible to claim that high-quality data can be collected on CloudResearch but not on Mechanical Turk. Researchers and other users of MTurk can also develop tools that address data quality in behavioral science studies. In other words, as we stress throughout this paper, a research platform cannot be evaluated in the absence of the tools that are available to use on that platform.

Our results present clear evidence of the high data quality researchers can gather from CloudResearch relative to other platforms. Nevertheless, the arguments made by Peer et al still

have the potential to distort how researchers and the people responsible for making decisions about research funding think about evaluating online participant platforms. Specifically, Peer et al advance two ideas about online participant platforms that we believe are mistaken. First, the authors make several statements that encourage a fixed view of data collection platforms, and second they ignore that platforms often have complementary strengths and weaknesses.

On the first issue, Peer et al. claimed that their goal was to go "beyond the basic question of whether an online platform or panel is suitable for research, to the more advanced question of *which* online platform or panel can produce the *best* data quality, and on which aspects the platforms and panels differ" (pg. 3). Elsewhere the authors suggest that researchers may substitute a platform-level reputation for a participant-level reputation or the use of attention checks. And finally, the authors present sample-level point estimates for things like demographics as evidence of platform-wide characteristics, ignoring the fact that sampling methods affect the generalizability of these point estimates.

Contrary to the view offered by Peer et al., online data collection platforms are dynamic marketplaces where things change. For evidence of this, one need look no further than Mechanical Turk. From 2010 to 2018 researchers published scores of papers investigating data quality on MTurk and outlining best practices for maintaining data quality. In general, these papers demonstrated that MTurk was a place where researchers could gather very high quality data (Buhrmester et al., 2011; Crump et al., 2013; Hauser & Schwartz, 2016; Mullinix et al., 2015). But then, in 2018 some threats to data quality emerged on MTurk (Moss & Litman, 2018), resulting in a substantive drop in data quality. As the data we present here demonstrate, CloudResearch has developed tools that circumnavigate these threats. Thus, the data quality of a platform can change over time and platforms may need to develop and implement novel tools to

address emerging threats to data quality. To insinuate, as done by Peer et al, that any particular platform is protected from online fraud based on one specific demonstration of satisfactory data quality is erroneous.

On the second issue about the strengths and weaknesses of different platforms, we believe Peer et al. advance a view that is much too narrow about the utility of different participant recruitment platforms. While it is generally true that sites like Mechanical Turk and Prolific have been found to have more engaged participants that produce higher quality data than market research panels (Chandler et al., 2019; Kees et al., 2017), the size and reach of market research panels makes them too valuable of a resource to dismiss based on a sample of mixed data quality.

For example, platforms like MTurk and Prolific provide access to tens of thousands of participants in any given month (Robinson et al., 2019), whereas market research panels provide access to tens of millions of people worldwide (Litman, Robinson, & Rosenzweig, 2020). The size of market research panels makes it possible to run a wide variety of studies that are simply not possible with smaller platforms. As just one example, *The New York Times* contracted with Dynata—one of the panels considered in Peer et al—in July 2020 to gather 250,000 survey responses from people all over the U.S. in just a two-week period (Katz, Sanger-Katz, Quealy, 2020). This survey asked people about their mask-wearing and other health-related behaviors during the COVID-19 pandemic. Not only was this study's sample size over a two week period larger than all the active MTurk and Prolific participants annually, but this study and others like it (e.g., Litman et al., 2020) demonstrate geographic targeting capabilities that behavioral scientists can't replicate when using smaller platforms such as MTurk or Prolific. Surely, these

studies are valuable contributions to the understanding of human behavior and not a misuse of research funds.

Beyond the benefits of running studies that cannot be conducted on smaller platforms, researchers can employ methodological techniques that improve the quality of data from market research panels (e.g., Chandler et al., 2019). By screening inattentive or unengaged participants out of a study, researchers can maintain the benefits of a large and diverse participant pool without incurring the costs of discarding numerous participants.

Elsewhere, we have advocated for a "fit for purpose" framework when choosing online participant recruitment platforms (Litman, Robinson, & Rosenzweig, 2020). This approach advocates for the selection of a participant platform based on the best match between a study's goals and the platform's strengths and weaknesses. For example, platforms like CloudResearch, Prolific, and MTurk are much more likely to have access to respondents who are willing to stay engaged in very long studies (> 30 minutes). Therefore, when conducting a long study that requires high participant engagement these platforms would have a high fit for the study's purpose. However, these platforms would not have a high fit for a study seeking to recruit participants from a specific U.S. zip code or a study seeking to sample demographically hard-to-reach groups. Ultimately, it is this complementary view of online platforms that allows researchers to conduct the widest range of online studies and to benefit from the range of resources available among online platforms.

References

Blais, A. R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making,*1,(1).

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,* 3–5. https://doi.org/10.1177/1745691610393980

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

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,* 2022-2038. https://doi.org/10.3758/s13428-019-01273-7

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.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48,* 400-407.

Katz, J., Sanger-Katz, M., & Quealy, K., (2020, July 17). A detailed map of who is wearing masks in the U.S. *The New York Times.* Retrieved from http://www.nytimes.com

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46,* 141–155. https://doi.org/10.1080/00913367.2016.1269304

Litman, L., Hartman, R., Jaffe, S. N., & Robinson, J. (2020, July 24). County-level recruitment in online samples: Applications to COVID-19 and beyond. *PsyArXiv*. https://doi.org/10.31234/osf.io/g3xw7

Litman, L., Robinson, J. & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavioral Research Methods, 49,* 433–442. https://doi.org/10.3758/s13428-016-0727-z

Litman, L., Robinson, J., & Rosenzweig, C. (2020). Beyond Mechanical Turk: Using online market research platforms. In L. Litman and J. Robinson (Eds.) *Conducting Online Research on Amazon Mechanical Turk and Beyond* (217-233). Sage Academic Publishing. Thousand Oaks: CA

Litman, L., Rosenzweig, C., & Moss, A. J. (2020, July 15). New solutions dramatically improve research data quality on MTurk. *CloudResearch blog.* https://www.cloudresearch.com/resources/blog/page/4/

Mazar N, Amir O, Ariely D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research,* 45, 633-644. doi:10.1509/jmkr.45.6.633

Moss, A. J., & Litman, L. (2018, September, 18). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. *CloudResearch blog.* https://www.cloudresearch.com/resources/blog/page/9/

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2,* 109–138. https://doi.org/10.1017/XPS.2015.19

Peer, E., Rothschild, D. M., Evernden, Z., Gordon, A., & Damer, E. (2021). *MTurk, Prolific or panels? Choosing the right audience for online research*. Available at SSRN: https://ssrn.com/abstract=

Peer, E., Vosgerau, J. & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavioral Research Methods, 46,* 1023–1031 (2014). https://doi.org/10.3758/s13428-013-0434-y

Robinson, J., Rosenzweig, C., Moss, A.J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLoS ONE 14(12)*: e0226394. https://doi.org/10.1371/journal.pone.0226394