




The Burden for High-Quality Online Data Collection Lies With Researchers, Not Recruitment Platforms

Christine Cuskley¹ and Justin Sulik² 

¹Language Evolution, Acquisition, and Development Group, Newcastle University, and

²Cognition, Values and Behavior, LMU Munich

Abstract

A recent article in *Perspectives on Psychological Science* (Webb & Tangney, 2022) reported a study in which just 2.6% of participants recruited on Amazon’s Mechanical Turk (MTurk) were deemed “valid.” The authors highlighted some well-established limitations of MTurk, but their central claims—that MTurk is “too good to be true” and that it captured “only 14 human beings . . . [out of] $N = 529$ ”—are radically misleading, yet have been repeated widely. This commentary aims to (a) correct the record (i.e., by showing that Webb and Tangney’s approach to data collection led to unusually low data quality) and (b) offer a shift in perspective for running high-quality studies online. Negative attitudes toward MTurk sometimes reflect a fundamental misunderstanding of what the platform offers and how it should be used in research. Beyond pointing to research that details strategies for effective design and recruitment on MTurk, we stress that MTurk is not suitable for every study. Effective use requires specific expertise and design considerations. Like all tools used in research—from advanced hardware to specialist software—the tool itself places constraints on what one should use it for. Ultimately, high-quality data is the responsibility of the researcher, not the crowdsourcing platform.

Keywords

online research, crowdsourcing, Mechanical Turk, data quality, attention checks, ethics

Amazon’s Mechanical Turk (MTurk) is a crowdsourcing platform used widely across the social and behavioral sciences (Anderson et al., 2019; Buhrmester et al., 2018; Crump et al., 2013; Mason & Suri, 2012; Paolacci & Chandler, 2014; Paolacci et al., 2010; Zallot et al., 2021). Webb and Tangney (2022; henceforth W&T) described a study conducted on MTurk involving a mental-health questionnaire.¹ Their view is that MTurk is “too good to be true” because of “bots and bad data” and that only 2.6% of their total sample (14 of $N = 529$) were “human beings.” W&T admit that their report is “not an empirical assessment of validity of all MTurk data” (p. 4), but it has been received as implying just that. Discussion of the study (for instance, see tweets at <https://sage.altmetric.com/details/138168456>) has interpreted the specific “2.6%” statistic as applying to MTurk in general, interpreting W&T’s conclusion as “a cautionary tale on the value of MTurk samples” or a “warning call” against MTurk.

W&T identify a core issue as “ambiguity” and “opacity” surrounding MTurk data quality. Despite the unique

challenges of online research,² there are effective strategies researchers can employ to ensure data quality. Although these strategies are documented in a substantial literature, this does not mean they are straightforward to understand or implement, or that they are unchanging; the best practice for any method—like science itself—evolves. Like other tools in psychological science, the effective use of MTurk or other online crowdsourcing tools is complex. MTurk is only “too good to be true” if researchers have misconceptions about the considerable labor involved in collecting high-quality data online or the time and effort needed to develop the requisite expertise. The net costs of using MTurk ethically and rigorously are not necessarily lower than using in-person studies (and may sometimes be more), but in any case, the costs are different: The costs of up-front investment in experimental design and

Corresponding Author:

Justin Sulik, LMU Munich, Cognition, Values and Behavior
 Email: justin.sulik@gmail.com

data-quality monitoring counterbalance the benefits of quicker data collection and access to a more diverse participant pool (Moss et al., 2020; Rodd, 2023).

In other words, there is a substantial burden associated with collecting high-quality behavioral data online; however, we argue that this burden lies on researchers, not recruitment platforms. We should not expect these platforms—which are involved in data collection for a wide range of sectors—to have interests that align completely with those of academic researchers, and we should want to maintain rigorous control over our own data quality as part of a comprehensive research pipeline.

We aim to explain why the estimated 2.6% participant viability in W&T's study is unusually low and how some claims in W&T's article are indicative of widespread misunderstanding of what MTurk is or how to use it. W&T mention some strategies drawn from organizational psychology to ensure data quality (Keith et al., 2017) but not the vast literature on using MTurk effectively, including specific guidance for the focus of their study: mental health and clinical psychology (Agle et al., 2022; Chandler et al., 2020; Ophir et al., 2020; Zorowitz et al., 2023). Our discussion references the robust and growing literature on MTurk, but this is not a tutorial on how to run a study online—that would require considerable depth (data quality has various complex roots in design and recruitment strategies) and breadth (design and recruitment must adapt to fit a specific study's needs). This has received book-length treatment elsewhere (e.g., Litman & Robinson, 2020), and there are many shorter guides to crowdsourcing (Bauer et al., 2020; Hauser et al., 2019; Rodd, 2023; Zallot et al., 2021). Instead, we offer mental models of MTurk (or crowdsourcing platforms in general) to guide researchers' expectations.

Too Good to Be True?

W&T assessed their 529 respondents on six criteria, with the number of respondents lost to each criterion in parentheses:³ (a) eligibility (−193); (b) performance on a consent quiz (−136); (c) noncompletion (−60; 44 abandoned the task, 16 clicked through directly to payment); (d) failing attention checks (−16); (e) response time (−47); and (f) unusual answers to open-ended questions such as “Who are you? Write ten sentences below describing yourself as you are today” (−77).

W&T frame their 2.6% claim in terms of validity, but *validity* typically refers to properties of a measure (e.g., whether it captures the phenomenon of interest), often evaluated alongside *reliability* (e.g., the accuracy of a measure across contexts). As validity is generally not a property of participants, we refer instead to those who

were “ineligible” (and did not participate) and those who were “unviable” (failing data-quality checks). Second, we highlight elements of the study design that either depart from established best practice in online research or stem from issues that likely go beyond online data collection.

Ineligible respondents are not participants

The majority of the reported participant sample ($N = 529$) did not complete the task (329 from criteria “a” and “b”; 44 from criterion “c”). But these “participants” did not actually participate. Participants who do not complete a study (in the lab or online) should be reported as having voluntarily withdrawn or as having been excluded.⁴ Exclusions based on demographic characteristics or understanding of informed consent are reasonable, but such exclusions are fundamentally not part of the sample.

Beyond demographic segmenting provided by MTurk—part of criterion “a”—there are various ways to perform eligibility exclusions, though these involve familiarity with web design (Hauser et al., 2019), multiphase recruitment (Hydock, 2018; Springer et al., 2016), or third-party services that prescreen MTurk workers (or “Turkers”; e.g., CloudResearch offers fine-grained age filters and checks whether reported demographics are accurate). All of these cost time and money. Such costs must be considered when choosing to crowdsource data; doing so would have averted most of the issues W&T encountered.

It may seem alarming that the consent quiz in criterion “b” had a 40% failure rate (136/336 remaining after criterion “a”). However, participants do not read consent forms carefully in in-person studies, either (Douglas et al., 2021). W&T's rate is in range for in-person studies—from ~25% for simple components of informed consent to ~45% for more complex components (Tam et al., 2015). It may be a general problem for participant-based research that participants do not carefully read consent information, but it is not a problem specific to online data collection or MTurk.

Good data is not a given

The attention checks in criterion “d” (known as *instructional manipulation checks*, or IMCs) are commonly used to assess data quality and participant viability both online and in the lab. W&T imply that unviable participants were not human (i.e., bots), but humans also produce unviable data during in-person studies (Hauser & Schwarz, 2016; Necka et al., 2016). For W&T,

16 participants failed these checks (~10% of the 156 remaining). In-person studies have IMC failure rates in the range of 14 to 18% for motivated or monitored participants and up to 28% for unmotivated participants (Oppenheimer et al., 2009). The 10% rate for W&T's criterion "d" does not reflect an MTurk-specific phenomenon,⁵ and worries about pervasive bots on MTurk have been debunked (Ahler et al., 2021; Kennedy et al., 2020; Moss et al., 2021).

Recent assays of quality on unfiltered MTurk find that ~60% of Turkers provide acceptable quality data and ~40% unacceptable (Hauser et al., 2022), though more may display "careless behavior" in a broader, less pernicious sense, highlighting the importance of nuanced approaches to data quality (Brühlmann et al., 2020). The aforementioned filtering services, such as CloudResearch, decrease the rate of unviable responses substantially (Douglas et al., 2023; Hauser et al., 2022). In our experience, combining such services with two-stage recruitment (in which researchers rerecruit participants who had previously passed attention checks) further reduces unviability by an order of magnitude. Again, such strategies cost time or money, and researchers choosing to recruit online must accept such costs if they want good data.

However, IMCs reflect only localized attention and not global task attention (Gummer et al., 2021) and are only moderately effective in bot detection (Pei et al., 2020; Storozuk et al., 2020); they also have measurement problems (Hauser et al., 2019) and frustrate participants (Silber et al., 2022). We thus advise against treating IMCs as straightforward indices of attention, in view of the fact that even attentive participants might still fail an occasional item. Our own convention is to count participants as unviable only if they fail more than one attention-check item. However, we recommend prioritizing more sophisticated techniques than IMCs. These track patterns of responses within and between participants (Buchanan & Scofield, 2018; Curran, 2016; Dupuis et al., 2019; Wood et al., 2017; see SM10 in Sulik et al., 2023, for a worked example, available at <https://osf.io/xw23p>).

Raw completion times are not diagnostic

Overall task-completion times (criterion "e") are not informative gauges of data quality. More informative indicators would be rates (e.g., hourly-equivalent participant-payment rates or seconds taken per response) and nuanced tracking of response behavior.

W&T's reported rate of viability is likely affected by the compensation offered and the length of the task. Assuming their reference to "minimum wage" means

the U.S. federal minimum wage (\$7.25/hour), and given the reported upper estimate of completion time (50 min), W&T presumably offered ~\$6 as compensation. Minimum wage in most U.S. states and territories is higher than the federal minimum, and tools used by Turkers to track compensation rates (e.g., <https://turker-view.com/>) consider the federal rate to be low pay. Our own policy is to offer a minimum \$12/hr equivalent; we arrived at this rate after considering mean response times from pilots with MTurk participants, calculating time from when participants consent and begin the survey, not from when they open the survey window. We urge against making assumptions about respondents' likely behavior and against estimating completion times using non-MTurk samples (or samples not from the recruitment platform the study will use).

The compensation offered by W&T was not only relatively low, but the task itself was very long. Online surveys should ideally be 5 to 15 min (Aguinis et al., 2021; Moss et al., 2023; Revilla & Höhne, 2020), with ~20 min being a reasonable maximum (Cape & Phillips, 2015; Chandler, 2023; Revilla & Ochoa, 2017). It would come as no surprise if many of W&T's participants abandoned the task or rushed to get a better hourly wage or avoid boredom (indeed, the lower bound of their reported completion times aligns with the maximum recommended time for a survey noted above). W&T mention that "owing to the compensation structure, 'workers' have little incentive to invest the extra time and thought required by open-ended qualitative items" (p. 4). However, this overlooks the fact that researchers, not MTurk, choose the compensation structure and task duration.

Motivation and incentives for online studies differ from in-person studies (Hauser et al., 2019). There is a complex and evolving relationship between compensation and data quality online: U.S. Turkers identify money as a primary motivator and are less likely to accept low-paying tasks (Litman et al., 2015). Additionally, low pay leads to participant frustration (Fowler et al., 2022), and as the results from Litman et al. (2015) involved a short 6-min task, such issues may be compounded in long questionnaires. Participants speed up toward the end of longer⁶ questionnaires; they are also less likely to move sliders, they give shorter responses to open-text questions, and they grow increasingly careless (Bowling et al., 2022; Cape & Phillips, 2015).

Researchers (and ethical-review bodies) considering MTurk should be aware of the well-documented potential for worker exploitation (Pittman & Sheehan, 2016), adjusting compensation to scale with what they would expect to pay in the lab. Further, researchers should carefully and honestly communicate their expectations about

time or attention to participants (Galesic & Bosnjak, 2009; Hauser et al., 2019; Zallot et al., 2021). If MTurk is to be considered “cheap” (a characterization we resist in this simple form), it is not because the pay rate should be lower, but because researchers can get data from MTurk for short studies; it may be difficult to get participants into the lab for similarly short periods.

Overall, completion times are only a useful index of data quality in the context of other information: A two-hour completion time could reflect a participant who is working distractedly while watching TV, but it could just as easily represent someone who opens the survey window and does not immediately begin the survey, but responds attentively once the survey is launched. In our experience, the latter scenario is common. To distinguish such cases, researchers can record times between clicks, track mouse movements, or monitor when a participant has moved away from the survey’s browser window. All of these can be achieved with flexible platforms such as jsPsych (De Leeuw, 2015), but again, this requires some skill development.

Open-ended questions must be motivated

Finally, W&T used two open-ended questions⁷ for which they sought 10 sentences in response and at least 20 sentences total (criterion “f”). Participants were eliminated if their responses were deemed unusual, nonsense, contradictory, or repetitive across participants. Assuming that this approach is a variant of the 20-statements test (Kuhn & McPartland, 1954), there is no published comparison of performance on this task online versus in the lab, nor are there established procedures for coding how humans (as opposed to bots) respond to this. It is not clear, for example, that replicated statements like “I will . . . get married” or “I will . . . buy a car” across respondents indicate that respondents are not likely to be human or are providing low-quality data (given how these are likely future events for respondents in their early twenties).

A common complaint among Turkers is frustration or confusion with open-ended or repetitive questions (Fowler et al., 2022). This is less a matter of comprehending the question and more a matter of not understanding why the question is being posed or what researchers expect in terms of a satisfactory response. Participants may deliberately give unusual responses when the perceived pointlessness of such questions inspires frustration (Fowler et al., 2022).

It is thus difficult to determine the root cause of the poor-quality data obtained in W&T’s study. Odd responses may have been generated by alleged bots, or they may be merely the last gasp of thoroughly bored participants. Because W&T did not filter by location,⁸ there is also the possibility of a language barrier.

Understanding MTurk

Well-recompensed studies that rigorously check multiple interrelated indices of data quality show a radically different picture of data quality on MTurk than W&T suggest (Hauser et al., 2022). Their reported 2.6% rate is off by more than an order of magnitude relative to even the most conservative estimates (Brühlmann et al., 2020; Chmielewski & Kucker, 2020; Douglas et al., 2023). Although W&T’s experience does not reflect the rates of data quality that researchers should expect from MTurk, it highlights a larger problem: many researchers seem to be operating with the wrong mental model of what MTurk is and how it should be used. Disappointing results may reflect mismatches between many researchers’ expectations and what MTurk actually provides.

Expect nothing from MTurk in isolation

Researchers should proceed as if MTurk offered no promise of data quality. MTurk is used across heterogeneous branches of academic research, and even this is a small share of their business: Private-sector companies use it for disparate tasks (Schmidt, 2015) with diverse criteria. Reporting data-quality metrics that span all these areas is not feasible, so researchers should consult recent estimates for specific fields (e.g., psychology, Douglas et al., 2023; advertising, Berry et al., 2022; political science, Kennedy et al., 2020) and calibrate their expectations accordingly.

The opportunity costs involved in ensuring data quality on MTurk have led to the emergence of crowdsourcing platforms, such as Prolific, that were originally designed for academic research. Although some researchers may prefer such purpose-built platforms, Prolific is not immune to the data-quality issues inherent to crowdsourced approaches (see Charalambides, 2021), and using CloudResearch to filter MTurk participation results in similar—and sometimes even higher—quality than Prolific (Douglas et al., 2023; Peer et al., 2022).

Researchers should think of recruitment platforms like MTurk as the classifieds section of a newspaper: It is just a place to list or browse offerings. Recruiting directly from MTurk is like offering a job to the first N people who respond to a classified ad, whereas paying for CloudResearch-filtered participants is like engaging a recruitment agency. Alternatively, researchers could take the time and effort to vet their own participants using multistage recruitment with multiple mutually corroborative indices of quality (Storozuk et al., 2020).

Expect little from basic filtering

MTurk’s main signal of participant quality is the HIT⁹ Approval Rate (HAR), tracking the percentage of a

participant's previous work that was approved rather than rejected by requesters. Various papers on using MTurk for academic studies recommend filtering out workers with an HAR below some high threshold (commonly > 95%, as in W&T). However, high HAR is not a guarantee of good data quality, even though low HAR may indicate a likelihood of low-quality data (Ahler et al., 2021; Hauser et al., 2022; Peer et al., 2022). Conversely, MTurk's most prolific participants typically have a high HAR, so researchers can access more naive populations if they set a lower threshold (Robinson et al., 2019). How (or whether) a researcher wants to use a worker's HAR as a filter during recruitment will depend on specific study requirements.

However, researchers should focus on the broader process of recruiting online, rather than just checking boxes or setting filters such as this. Whether in person or online, it remains a researcher's responsibility to develop a rigorous recruitment strategy that is tailored to a particular study's needs (Chandler, 2023; Rodd, 2023). Online researchers must check who they are recruiting, build data-quality indices into their studies from the ground up, and demonstrate a rigorous approach to data quality for their own sake and for their audience's sake (whether editors, reviewers or readers).

Do not expect MTurk to be both cheap and fast

Ever since the first appearance of the term "web experiment" in psychology (Reips, 2000), there have been widespread assumptions that online data collection is quick and cheap, or simply a matter of translating an in-lab study for the web browser (Rodd, 2023). Such assumptions persist despite decades of evidence to the contrary. This perception of low cost may be a holdover from the early MTurk "boom": Low per-participant costs were considered a main draw, but concerns about exploitation mean this is no longer the default (Pittman & Sheehan, 2016). As with in-person studies, it remains the ethical responsibility of researchers to pay participants at a rate that is neither exploitative or coercive.

MTurk is an efficient way to collect data once a study has been effectively designed, but that does not mean that researchers should expect overall costs to be lower; in fact, efficient recruitment means spending extra resources on other aspects of study design. Whether vetting their own participants or paying third-party filter services, researchers should harness the efficiency of MTurk to iterate over multiple versions of their study, benchmarking comprehension and attention, and using this to fine-tune their instructions and response formats.

Although W&T reported running a pilot of their study, it is not clear whether they did so on MTurk. If

they had done so with 30 participants, and if their exclusion criteria worked as described, a pilot would have yielded, at most, one viable participant by their standards (and possibly none). In that case, recruitment of the full sample should not have proceeded until the entire approach was reconsidered. We stress that reducing unviable responses to 0 is unlikely, but a pilot study should aim for a low rate (say, 7%–10%) that suits a researcher's balance between available time and money. Targeted piloting counteracts the speed of recruitment on MTurk relative to in-person recruitment. Although the relative costs of different parts of the research pipeline may vary, rigorous behavioral research has a fairly fixed cost when considered holistically.

Expect to spend time developing technical skills and keeping them up to date

Many methods in psychological science require specific expertise, and studies employing them must be adapted to their affordances. There is no reason to think crowdsourcing data should be a lone exception. An eye-tracking study run without calibration of the device and without data preprocessing would yield misleading conclusions, and yet failure to do so would not undermine the value of eye trackers for psychology research more generally. Tools for studying complex problems evolve rapidly, but the pace of these changes does not absolve poor practice. High-tech landscapes have an even faster rate of change (cf. the sudden explosion of ChatGPT in recent years), so researchers opting for online research should expect to keep abreast of current developments.

Relevant technical skills include lessons from web development and data science (Hauser et al., 2019). Web applications generally undergo extensive testing and revision before launch, and data wrangling, cleaning, and preprocessing are major parts of any data-science pipeline (Wickham & Grolemund, 2016). Specific skills for online research include coding for the web (e.g., in JavaScript), which allows one to enhance the relatively constrained affordances of survey platforms such as Qualtrics or to benefit from the flexibility of custom libraries such as jsPsych (De Leeuw, 2015). Other platforms lie between these extremes (e.g., Gorilla Psych, <https://gorilla.sc>; Labvanced, <https://labvanced.com>). Even without coding, basic attention to the control flow of a task would prevent participants clicking through to the end of the study (cf. W&T's criterion "d").

If a researcher cannot invest time in skill development, there is the option to pay expert consultants (e.g., offered by both Gorilla and CloudResearch). Whether the investment comes in the form of time or other

resources, the net cost of conducting effective research online is substantial.

Expect to prioritize participant perspectives

Some researchers may think of online studies as being like in-person laboratory studies occurring at a distant undisclosed location (Rodd, 2023). But online research benefits from considering online participants' perspectives throughout the design and recruitment process. Our emphasis on expertise notwithstanding, everyone has to start somewhere. While researchers are working to build expertise, they can benefit immensely from the perspectives of Turkers and their communities (Fowler et al., 2022; Schmidt, 2015; Silber et al., 2022), including browsing forums where Turkers discuss their experiences (e.g., <https://www.reddit.com/r/TurkerNation/>, <https://turkopticon.net>, <https://turkerview.com>). Even more simply, while iteratively piloting a new study, one should regularly ask for participant feedback (Aguinis et al., 2021) and act on it, offering bonus payments where such feedback is useful. Indeed, we consider that every crowdsourced study should end by asking for open-ended user feedback, and it is relatively easy to build a pool of trusted testers.

The issue extends beyond basic survey-design principles. For instance, boredom is a major issue on MTurk (Fowler et al., 2022), though often not considered in the lab. Researchers should accept that they are competing with everything else on the Internet when it comes to people's attention, and they should design engaging studies accordingly. MTurk is best suited to short studies with broad participant requirements and without dozens of Likert-style questions (aptly called "bubble hell" by workers; Fowler et al., 2022).

Our emphasis on technical skills and perspective-taking converge in UX (user experience in the context of web development). This includes designing interfaces for ease of use, taking into account user perceptions and motivations, and it forms a natural connection with gamification to enhance participant motivation in online studies (Rodd, 2023; Tinati et al., 2017). Ultimately, task design should make it effortless for respondents to give the kinds of responses researchers need, but effortful for them to try to bypass researcher needs. Even then, not every survey or experiment is suitable for online recruitment. Just as not every research question could be effectively addressed using eye tracking.

Conclusions

Webb and Tangney (2022) reported a study with alarmingly low rates of data quality using MTurk. We used their report to highlight how researcher expectations of

MTurk often don't match the affordances of the platform, though these challenges are well documented in the literature. MTurk is a large marketplace where tasks can be posted for workers to complete, giving researchers access to a diverse participant pool (Buhrmester et al., 2011; Smith et al., 2015) and relatively quick data collection once a task has been effectively designed. Alone, it is not effective as a filter for task eligibility or participant viability. The responsibility of ensuring a desirable sample and high-quality data still lies with the researcher, requiring either careful design and specialized skills or the hiring of services or experts. Information and strategies for how to accomplish this using MTurk are available in literature going back over a decade, much of which includes periodic estimates of data quality and how this changes over time, alongside task- and discipline-specific recommendations. This literature should form the basis of a researcher's assessment of whether the costs and benefits inherent to the platform are a fit for their research goals. MTurk may not always be an appropriate tool, given the skills of the research team or the nature of the research question. It is only "too good to be true" if researchers assume, despite a growing and robust literature to the contrary, that it is a magic bullet that will drastically reduce the overall costs of human-subjects research.

Transparency


Action Editor: Rogier Kievit

Editor: Interim Editorial Panel

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Justin Sulik  <https://orcid.org/0000-0002-0978-9496>

Notes

1. Webb and Tangney did not report detailed methods, but rather only certain aspects of the study relevant to excluding participants; we rely here only on what they reported.
2. For tabular summaries of challenges in online research, including how it may differ from in-person studies or how some perceptions of difference are in fact unfounded, see Aguinis et al. (2021, especially their Table 2); Necka et al. (2016, Table 2); Hauser et al. (2019, Table 1); Thomas and Clifford (2017, Table 1); Lowry et al. (2016, Table 1); and Lu et al. (2022, Table 1).
3. Note that the criteria here (and their lettering) reflect the order in which W&T described these criteria being applied in detail (from p. 2), but their description of their screening process (p. 1) puts attention checks as "c" and completion as "d."
4. We wish to emphasize that although participants who withdraw voluntarily are indeed participants and should be reported as such (though researchers must consider whether this compromises random assignment to conditions; Zhou & Fishbach,

2016), this is distinct from respondents to a recruitment call who do not meet eligibility criteria, and thus do not participate. To phrase this in terms of in-lab recruitment, we would not consider a bilingual French-English speaker who responded via email to a recruitment call for monolingual English speakers (and thus was not invited to participate) as part of the participant sample for the study.

5. We note that this rate is conditional on participants having passed earlier exclusion criteria and thus does not represent how likely it is that MTurk participants in general would fail such checks. W&T do not report information about respondents who may have failed multiple criteria. Open sharing of raw data would help clarify such issues.

6. Even then, “long” in Cape and Phillips (2015) means ~20 min.
7. Given the current availability of large language models (e.g., ChatGPT), it is now even more difficult to ensure the quality of open-ended text responses (Veselovsky et al., 2023). These tools were released in late 2022 and would not have been available at the time of W&T’s data collection. Nonetheless, even before widespread use of convincing text-generation tools, it was not considered best practices to use burdensome open-ended questions in online studies (Crawford et al., 2001; Fowler et al., 2022; Liu & Wronski, 2018; though Aguinis et al., 2021, disagreed). Unless one’s study is specifically about the human ability to generate or summarize text (such a study is probably best suited for the laboratory), researchers can mitigate the impact of AI tools by communicating better with participants and by developing (non-AI) technical skills, rather than engaging in a high-tech arms race with AIs. For instance, these skills include tracking or blocking copy+paste events in the webpage code (Veselovsky et al., 2023).

8. W&T noted that they were unable to filter on IP address or a more specific location because of IRB restrictions on collecting identifying information. They also noted that filtering by IP address is ineffective regardless (Dennis et al., 2020); however, MTurk also offers location filtering at the state level based on participants’ self-report of their location; this allows requesters to confine their sample further without collecting information about a specific location. Location filtering is further improved by services such as CloudResearch. In general, filtering in real time on the basis of IP address or other specific location information does not require collecting or recording this information (e.g., with back-end scripting) and should not be a barrier in ethical review.
9. HIT (human intelligence task) is an MTurk-specific term for the tasks posted online for MTurk workers to do.

References

- Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon’s Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavioral Research Methods*, *54*, 885–897. <https://doi.org/10.3758/s13428-021-01665-8>
- Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837.
- Ahler, D. J., Roush, C. E., & Sood, G. (2021). The micro-task market for lemons: Data quality on Amazon’s Mechanical Turk. *Political Science Research and Methods*, 1–20. <https://doi.org/10.1017/psrm.2021.57>
- Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850.
- Bauer, B., Larsen, K. L., Caulfield, N., Elder, D., Jordan, S., & Capron, D. (2020, November 14). *Review of best practice recommendations for ensuring high quality data with Amazon’s Mechanical Turk*. <https://doi.org/10.31234/osf.io/m78sf>
- Berry, C., Kees, J., & Burton, S. (2022). Drivers of data quality in advertising research: Differences across MTurk and professional panel samples. *Journal of Advertising*, *51*(4), 515–529.
- Bowling, N. A., Gibson, A. M., & DeSimone, J. A. (2022). Stop with the questions already! Does data quality suffer for scales positioned near the end of a lengthy questionnaire? *Journal of Business and Psychology*, *37*(5), 1099–1116.
- Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, Article 100022.
- Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*, 2586–2596.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.
- Buhrmester, M. D., Talaifar, 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.
- Cape, P., & Phillips, K. (2015). *Questionnaire length and fatigue effects: The latest thinking and practical solutions* [White paper]. Survey Sampling International. <https://silos.tips/download/white-paper-questionnaire-length-and-fatigue-effects-the-latest-thinking-and-pra>
- Chandler, J. (2023). Participant recruitment. In A. L. Nichols & J. Edlund (Eds.), *Cambridge handbook of research methods and statistics for the social and behavioral sciences* (pp. 179–201). Cambridge University Press.
- Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, *129*(1), 49–55.
- Charalambides, N. (2021). *We recently went viral on TikTok - here’s what we learned*. Prolific Blog. <https://www.prolific.co/blog/we-recently-went-viral-on-tiktok-heres-what-we-learned>
- Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473.
- Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review*, *19*(2), 146–162.
- 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), Article e57410. <https://doi.org/10.1371/journal.pone.0057410>

- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19. <https://doi.org/10.1016/j.jesp.2015.07.006>
- De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*, 1–12.
- Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting, 32*(1), 119–134.
- Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE, 18*(3), Article e0279720. <https://doi.org/10.1371/journal.pone.0279720>
- Douglas, B. D., McGorray, E. L., & Ewell, P. J. (2021). Some researchers wear yellow pants, but even fewer participants read consent forms: Exploring and improving consent form reading in human subjects research. *Psychological Methods, 26*(1), 61–68. <https://doi.org/10.1037/met0000267>
- Dupuis, M., Meier, E., & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods, 51*, 2228–2237. <https://doi.org/10.3758/s13428-018-1103-y>
- Fowler, C., Jiao, J., & Pitts, M. (2022). Frustration and ennui among Amazon MTurk workers. *Behavior Research Methods, 55*, 3009–3025.
- Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349–360. <https://doi.org/10.1093/poq/nfp031>
- Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research, 50*(1), 238–264.
- Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2022). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*. Advance online publication. <https://doi.org/10.3758/s13428-022-01999-x>
- Hauser, D. J., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 319–337). Routledge.
- 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*(1), 400–407.
- Hydock, C. (2018). Assessing and overcoming participant dishonesty in online data collection. *Behavior Research Methods, 50*, 1563–1567.
- 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*, Article 1359. <https://doi.org/10.3389/fpsyg.2017.01359>
- Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods, 8*(4), 614–629.
- Kuhn, M. H., & McPartland, T. S. (1954). An empirical investigation of self-attitudes. *American Sociological Review, 19*(1), 68–76.
- Litman, L., & Robinson, J. (2020). *Conducting online research on Amazon Mechanical Turk and beyond*. Sage.
- Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods, 47*(2), 519–528.
- Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review, 36*(1), 116–124.
- Lowry, P. B., D'Arcy, J., Hammer, B., & Moody, G. D. (2016). “Cargo Cult” science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels. *The Journal of Strategic Information Systems, 25*(3), 232–240.
- Lu, L., Neale, N., Line, N. D., & Bonn, M. (2022). Improving data quality using Amazon Mechanical Turk through platform setup. *Cornell Hospitality Quarterly, 63*(2), 231–246.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23.
- Moss, A. J., Hauser, D. J., Rosenzweig, C., Jaffe, S., Robinson, J., & Litman, L. (2023). Using market-research panels for behavioral science: An overview and tutorial. *Advances in Methods and Practices in Psychological Science, 6*(2). <https://doi.org/10.1177/25152459221140388>
- Moss, A. J., Rosenzweig, C., Jaffe, S. N., Gautam, R., Robinson, J., & Litman, L. (2021, June 11). *Bots or inattentive humans? Identifying sources of low-quality data in online platforms*. <https://doi.org/10.31234/osf.io/wr8ds>
- Moss, A. J., Rosenzweig, C., Robinson, J., Jaffe, S. N., & Litman, L. (2020, April 28). *Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages*. <https://doi.org/10.31234/osf.io/jbc9d>
- Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLOS ONE, 11*(6), Article e0157732. <https://doi.org/10.1371/journal.pone.0157732>
- Ophir, Y., Sisso, I., Asterhan, C. S. C., Tikochinski, R., & Reichart, R. (2020). The Turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. *Clinical Psychological Science, 8*(1), 65–83.
- 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. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3), 184–188.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411–419.
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643–1662.
- Pei, W., Mayer, A., Tu, K., & Yue, C. (2020, April). Attention please: Your attention check questions in survey studies can be automatically answered. In *Proceedings of the International World Wide Web Conference 2020* (pp. 1182–1193). Association for Computing Machinery.
- Pittman, M., & Sheehan, K. (2016). Amazon's Mechanical Turk a digital sweatshop? Transparency and accountability in crowdsourced online research. *Journal of Media Ethics*, 31(4), 260–262.
- Reips, U. D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). Academic Press.
- Revilla, M., & Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, 62(5), 538–545.
- Revilla, M., & Ochoa, C. (2017). Ideal and maximum length for a web survey. *International Journal of Market Research*, 59(5), 557–565.
- 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), Article e0226394. <https://doi.org/10.1371/journal.pone.0226394>
- Rodd, J. M. (2023, April 13). *Moving experimental psychology online: How to maintain data quality when we can't see our participants*. <https://doi.org/10.31234/osf.io/2fhcb>
- Schmidt, G. B. (2015). Fifty days an MTurk worker: The social and motivational context for Amazon Mechanical Turk workers. *Industrial and Organizational Psychology*, 8(2), 165–171.
- Silber, H., Roßmann, J., & Gummer, T. (2022). The issue of non-compliance in attention check questions: False positives in instructed response items. *Field Methods*, 34(4), 346–360.
- Smith, N. A., Sabat, I. E., Martinez, L. R., Weaver, K., & Xu, S. (2015). A convenient solution: Using MTurk to sample from hard-to-reach populations. *Industrial and Organizational Psychology*, 8(2), 220–228.
- Springer, V. A., Martini, P. J., Lindsey, S. C., & Vezich, I. S. (2016). Practice-based considerations for using multi-stage survey design to reach special populations on Amazon's Mechanical Turk. *Survey Practice*, 9(5). <https://doi.org/10.29115/SP-2016-0029>
- Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*, 16(5), 472–481.
- Sulik, J., Ross, R. M., Balzan, R., & McKay, R. (2023). Delusion-like beliefs and data quality: Are classic cognitive biases artifacts of carelessness? *Journal of Psychopathology and Clinical Science*, 132(6), 749–760.
- Tam, N. T., Huy, N. T., Thoa, L. T. B., Long, N. P., Trang, N. T. H., Hirayama, K., & Karbwang, J. (2015). Participants' understanding of informed consent in clinical trials over three decades: Systematic review and meta-analysis. *Bulletin of the World Health Organization*, 93, 186–198.
- 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.
- Tinati, R., Luczak-Roesch, M., Simperl, E., & Hall, W. (2017). An investigation of player motivations in Eyewire, a gamified citizen science project. *Computers in Human Behavior*, 73, 527–540.
- Veselovsky, V., Ribeiro, M. H., Cozzolino, P., Gordon, A., Rothschild, D., & West, R. (2023, October 24). *Prevalence and prevention of large language model use in crowd work*. <https://doi.org/10.48550/arXiv.2310.15683>
- Webb, M. A., & Tangney, J. P. (2022). Too good to be true: Bots and bad data from Mechanical Turk. *Perspectives on Psychological Science*. Advance online publication. <https://doi.org/10.1177/17456916221120027>
- Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.
- Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, 8(4), 454–464.
- Zallot, C., Paolacci, G., Chandler, J., & Sisso, I. (2021). Crowdsourcing in observational and experimental research. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of computational social science* (Vol. 2, pp. 140–157). Routledge.
- Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4), 493–504.
- Zorowitz, S., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behavior and symptom measures. *Nature Human Behavior*. Advance online publication. <https://doi.org/10.1038/s41562-023-01640-7>