Perspectives on Psychological Science
1–4
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916221120027
www.psychologicalscience.org/PPS

Too Good to Be True: Bots and Bad Data From Mechanical Turk

Margaret A. Webb[1,2] and June P. Tangney[2]
[1]Department of Criminology, Max-Planck Institute for the Study of Crime, Security, and Law Freiburg im Breisgau; and [2]Department of Psychology, George Mason University

Abstract
Psychology is moving increasingly toward digital sources of data, with Amazon's Mechanical Turk (MTurk) at the forefront of that charge. In 2015, up to an estimated 45% of articles published in the top behavioral and social science journals included at least one study conducted on MTurk. In this article, I summarize my own experience with MTurk and how I deduced that my sample was—at best—only 2.6% valid, by my estimate. I share these results as a warning and call for caution. Recently, I conducted an online study via Amazon's MTurk, eager and excited to collect my own data for the first time as a doctoral student. What resulted has prompted me to write this as a warning: it is indeed too good to be true. This is a summary of how I determined that, at best, I had gathered valid data from 14 human beings—2.6% of my participant sample ($N = 529$).

Keywords
crowdsourcing, research data, data quality, Mechanical Turk

Amazon's Mechanical Turk (MTurk) and other crowdsourcing platforms have grown to be a dominant source of supposedly high-quality human-subjects research data (Buhrmester et al., 2018; Mortensen & Hughes, 2018). Today, in 2022, a Google Scholar search of "mechanical turk" returns 142,000 articles. In the behavioral and social sciences, researchers using these crowdsourcing platforms have been especially prolific: There are estimates that up to 45% of articles published in top psychology journals include at least one study conducted on MTurk (Zhou & Fishbach, 2016). Use of these platforms is anything but rare. With the COVID-19 pandemic in-person research became even more difficult, and the use of such crowdsourcing platforms expanded even further.

In response to concerns about MTurk data quality, certain filtering and checking techniques available within the platform and via back-end data cleaning have been increasingly recommended (Keith et al., 2017). To enhance quality in my own study, I followed these recommendations. On the platform, I filtered on location (United States) and Human Intelligence Task (HIT) approval rate (> 95% approval). I paid extra for MTurk premium filters for age of interest (18–25) and for English as a primary language, and I compensated participants at a minimum-wage rate. Of note, my study focused on suicidal ideation, so the Institutional Review Board required that we collect no identifying information (as is common for many clinical studies asking about sensitive topics). We were therefore unable to filter on IP address or geolocation. Importantly, both of these methods have been cited as ineffectual in preventing invalid responding (Dennis et al., 2020).

Feeling confident, on Monday, March 2, at 5:55 p.m., I launched my study and waited with bated breath for results to come in. Within 90 min, data collection was complete. I had 529 participants. I was ready to analyze, answer my question, and write!

The Screening Process

As the first step of analysis, I screened the 529 responses in the following order: (a) eligibility criteria, (b) performance on consent quiz, (c) performance on attention checks, (d) completion of the study, and (e) response time. Finally, as an additional and less-common check, I conducted (f) an examination of qualitative responses with the respondents who remained.

Corresponding Author:
Margaret A. Webb, George Mason University, Department of Psychology
Email: mwebb26@gmu.edu

Eligibility criteria

My study required participants to speak English and to be between the ages of 18 to 24 years. However, the standard premium MTurk filter of 18 to 25 years did not allow me to exactly capture this age range. In addition to the premium MTurk filters for age and language, I included screener questions at the beginning of the online survey asking for participants' age and primary language. Participants who did not meet criteria via these screener questions were unable to proceed with the study.

My additional screener questions caught seventy-one 25-year-olds (13.4% of the total sample), which was expected given that the premium filter allowed 25-year-olds to view the study. Alarmingly, despite the added-fee premium filter, an additional 118 participants (22%) reported ages of less than 18 years or more than 25 years—and some of these responses were clearly invalid (e.g., 0, 5, 100). Four participants (< 1%) did not meet criteria for the language requirement. This left 336 participants (64% of the original sample).

Consent quiz

Of the 336 participants who met eligibility criteria, 136 (40% of the remaining sample) twice failed a three-item true–false quiz regarding key information on the consent form (e.g., their right to end participation, their right to confidentiality, and researchers' ability to contact them). Participants were shown the informed-consent form and then given the quiz. If participants failed the quiz on their first try, they were shown the consent once again and then given a second opportunity to take the same quiz. If they failed on their second try, they were excluded from completing the study. If participants passed on the first attempt, they proceeded to the rest of the study. Just 200 participants remained (38% of the original sample) after the consent quiz.

Completion

Of the 200 participants who were eligible for the study and who passed the consent quiz, 60 (30% of the remaining sample) did not finish the 45-min survey. Sixteen (8%) of these participants clicked straight through to the end of the survey, without answering questions beyond the consent quiz, to obtain a payment code. At this point, 140 participants remained (26% of the original sample).

Attention checks

Three classic attention checks were included throughout the study at approximately the 25%, 50%, and 75% completion marks. The attention checks consisted of the following: (a) embedded on the Grit scale—"Select 'Somewhat like me' for this statement," (b) embedded on the Beck Depression Inventory—"1 – Select this option," and (c) embedded on the Borderline Personality Inventory—"Select 'Yes' for this statement."

Of the 140 participants who met eligibility criteria, passed the consent quiz, and completed all questions in the study, a total of 13 participants (9% of the remaining sample) failed one of the attention checks, with 3 participants (2% of the remaining sample) failing two or more. At this point in the study, 124 participants remained (23% of the original sample).

Unrealistic response time

The estimated completion time for the survey was 45 to 50 min. The distribution of response times among the participants still being considered for inclusion in the sample was heavily skewed. The completion time ranged from 4 min to 22 hr with a median of 27.99 min. We supposed that it might be possible that participants began the survey and then walked away from their computers to come back to it later in the day, or the next day. So as not to exclude participants for whom this may have been true, we did not include a ceiling for response times. Six participants took more than 17 hr to complete the survey. The longest completion time apart from these outliers was 3.2 hr.

From pilot tests prelaunch, we estimated that the survey would take 45 to 50 min. Given that many respondents on MTurk are highly familiar with these kinds of surveys, we supposed that they might be more adept at quickly completing surveys, and we used a conservative cutoff of 20 min to represent a realistic response time. Completion of the survey in less than 20 min is extremely improbable, so we considered this an indication of invalid responding. Of the 124 participants who passed attention checks, 47 participants (38% of the remaining sample) responded in less than 20 min and were excluded from the sample. At this stage, 77 participants remained (15% of the original sample).

At this point, we had eliminated 85% of our original sample using common techniques for screening out invalid responses. Already the proportion of invalid responses was extremely disheartening, and I might have just stopped there. But instead, given the qualitative components of the questionnaire and driven in part by morbid curiosity, I took a final step to examine qualitative responses.

Examination of qualitative responses

I evaluated the qualitative responses provided by the 77 participants who passed all five quantitative checks. Participants were asked: "Who are you? Write ten sentences below, describing yourself as you are today. (1) *I am . . .*

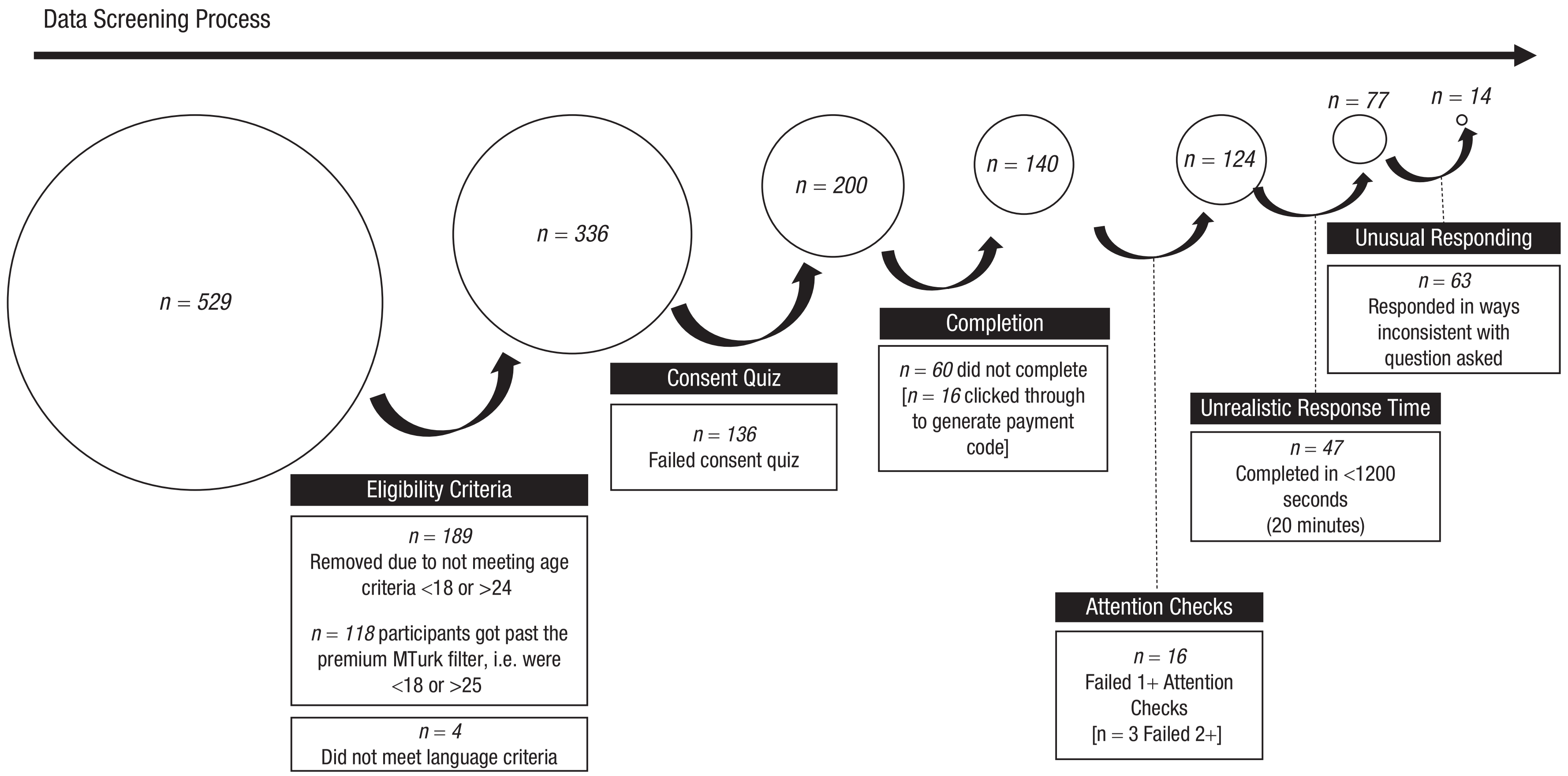
Data Screening Process
n = 529
n = 336
n = 200
n = 140
n = 124
n = 77
n = 14
Eligibility Criteria
n = 189
Removed due to not meeting age criteria <18 or >24
n = 118 participants got past the premium MTurk filter, i.e. were <18 or >25
n = 4
Did not meet language criteria
Consent Quiz
n = 136
Failed consent quiz
Completion
n = 60 did not complete [n = 16 clicked through to generate payment code]
Attention Checks
n = 16
Failed 1+ Attention Checks
[n = 3 Failed 2+]
Unrealistic Response Time
n = 47
Completed in <1200 seconds
(20 minutes)
Unusual Responding
n = 63
Responded in ways inconsistent with question asked

Fig. 1. Data-screening process. MTurk = Amazon Mechanical Turk.

[etc.]" and "Who will you be? Think about 1 week [1 year/10 years] from today. Write ten sentences below, describing yourself as you imagine you will be in 1 week [1 year/10 years]. (1) *I will be . . .* [etc.]" Responses were considered unusual if they were single words or phrases that did not make sense for the question asked (e.g., "relate your answer to the job," "soon") or were nonsense phrases (e.g., "fasty," "stress-Busting Lesiser Time"). In addition, responses within a participant's set that were contradictory (e.g., "a great man" and "a great woman") or response sets that were clearly replicated across numerous individual respondents (e.g., multiple respondents wrote "good personality," "lovable person," "get married," "buy a car," "build a new home," etc., in the same order) were also considered unusual and flagged as invalid. Of the 77 participants who passed all five quantitative checks, 63 participants (82% of the remaining sample) were flagged and removed for unusual responding. That left 14 participants (2.6% of the original sample).

Conclusion

After months spent on study formulation, costly participant payments, and hours of data cleaning, I was left with just 14 potentially real participants. For anyone, this would be a devastating outcome—but especially so for a graduate student with limited time and funds. I spent the next few days thinking about how to proceed and attempting to get reimbursed by Amazon (which, I am thankful to say, did return our funds in full after we shared our concerns about data quality). In reflecting on our outcome, my personal disappointment was eclipsed quickly by a sense of trepidation for what this means for science, which has come to rely so heavily on MTurk as a source of data. With approximately 15,000 articles published on MTurk in the first 6 months of 2022 alone, the ripple effects of bad MTurk data are enormous: failure to find replications, erroneous effects, lines of research based on false information. I feel compelled to write this as warning: If my 2.6% is even the lower bound of sample validity on MTurk, there is reason for skepticism and caution.

Epilogue

This article is not meant as an empirical assessment of the validity of all MTurk data; rather, it is an illustration of an individual experience. There is no way of knowing from these data alone what the true bound of validity is for all MTurk samples. For example, validity may vary depending on the length and nature of the survey. Respondents may pay more attention to brief surveys composed of close-ended questions, yielding more valid data. Owing to the compensation structure, "workers" have little incentive to invest the extra time and thought required by open-ended qualitative items, such as those included in our 45- to 50-min survey. Even so, this ambiguity is precisely the issue at hand. Our results paired with the overall opacity of MTurk's data quality leaves us with an unsettling and untenable uncertainty.

Transparency

Action Editor: Klaus Fiedler
Editor: Klaus Fiedler
Author Contributions
M. A. Webb: Conceptualization, Writing - original draft, Funding acquisition. J. P. Tangney: Writing - review & editing, Supervision. We chose to write this piece in the first person, as this is a reflection on the first author's lived experience and is not intended as an empirical assessment of the bound of validity for all MTurk samples. This manuscript did not undergo peer review but was accepted for publication after internal review by the editor and associate editors.
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.
Funding
The research reported in this manuscript was financially supported by a National Science Foundation Graduate Research Fellowship (Grant No. 2020291846) awarded to Margaret Webb. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

ORCID iD

Margaret Webb https://orcid.org/0000-0002-5552-0308

References

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.

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.

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

Mortensen, K., & Hughes, T. L. (2018). Comparing Amazon's Mechanical Turk platform to conventional data collection methods in the health and medical research literature. *Journal of General Internal Medicine*, *33*(4), 533–538.

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.