Empirical Article

Advances in Methods and Practices in Psychological Science
January-March 2026, Vol. 9, No. 1, pp. 1–12

DOI: 10.1177/25152459261424723
www.psychologicalscience.org/AMPPS

Chatbots Are Undermining Crowdsourced Research in the Behavioral Sciences: Detecting Artificial Intelligence–Assisted Cheating With a Keystroke-Based Tool

Michael W. Asher, Gillian Gold, Eason Chen, and Paulo F. Carvalho
Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania

Abstract

Generative artificial intelligence (AI) poses a significant threat to data integrity on crowdsourcing platforms, such as Prolific, which behavioral scientists widely rely on for data collection. Large language models (LLMs) allow users to generate fluent and relevant responses to open-ended questions, which can mask inattention and compromise experimental validity. To empirically estimate the prevalence of this behavior, we analyzed keystroke data from three studies ($N = 928$) on Prolific between May and July 2025. Using an embedded JavaScript tool, we flagged participants who pasted text or whose keystroke count was anomalously low compared with their response length. For each flagged participant, we manually compared detected keystrokes with their final response to determine if the text could have been typed. This confirmed that despite deterrence measures, approximately 9% of participants submitted responses consistent with AI assistance or other forms of outsourced responding. These participants outperformed noncheaters (by up to 1.5 *SD*), were more than twice as likely to share geolocations with other participants (suggesting possible proxy use), and exhibited lower internal consistency on questionnaire scales. Simulated power analyses indicate that this level of undetected cheating can diminish observed effect sizes by 10% and inflate required sample sizes by up to 30%. These findings highlight the urgent need for new detection methods, such as keystroke logging, which offers verifiable evidence of cheating that is difficult to obtain from manual review of LLM-generated text alone. As AI continues to evolve, maintaining data quality in crowdsourced research will require active monitoring, methodological adaptation, and communication between researchers and platforms.

Keywords

crowdsourced research, generative AI, cheating, automated responses, open data

Received 8/12/25; revision accepted 1/6/26

Online crowdsourcing platforms have become a staple in behavioral-sciences research. On websites such as Prolific, CloudResearch, and Amazon's Mechanical Turk (MTurk), researchers can quickly and affordably recruit participants. These platforms enable researchers to efficiently collect large data sets of relatively diverse participants (compared with traditional college-student samples; Buhrmester et al., 2011), contributing to a surge in online experiments across the behavioral sciences (Buhrmester et al., 2018). In 2022, more than 150,000 studies were published on Prolific alone (Tomczak et al., 2023), and several analyses suggest that at least half of all articles published in prominent social-psychology and cognitive-science journals include at least one crowdsourced study (Stewart et al., 2017; Zhou & Fishbach, 2016).

However, as crowdsourced research has grown in popularity, its use has been accompanied by increasing

Corresponding Author:
Michael W. Asher, Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania
Email: masher@andrew.cmu.edu

concerns about data quality, participant attentiveness, and the evolving threat of fraudulent and automated responses (Chmielewski & Kucker, 2020; Dennis et al., 2020; R. Kennedy et al., 2020; Peer et al., 2021; Webb & Tangney, 2024). These concerns have included automated or "bot" responses (often aided by virtual private networks [VPNs] or proxy servers to disguise locations and allow for operation of multiple accounts) flooding crowdsourced studies with multiple low-quality responses from the same individuals (Dennis et al., 2020). In online platforms, fraudulent or low-effort responses have been shown to reduce the validity of crowdsourced studies, introducing noise that can attenuate observed effect sizes and even create spurious associations (Chmielewski & Kucker, 2020; Wood et al., 2017).

Traditionally, open-ended questions have played a critical role in identifying low-quality, noncompliant participants. For example, researchers have flagged nonsensical responses, one-word responses unrelated to a prompt, or restatements of a question as clear indicators of inattention or inauthentic engagement (Chmielewski & Kucker, 2020; C. Kennedy et al., 2020; Price et al., 2024).

These strategies have proven effective at flagging low-quality submissions. For instance, Chmielewski and Kucker (2020) documented a sharp drop in data quality on MTurk beginning in 2018, which included a more than threefold increase in the number of responses that failed standard attention checks. However, by excluding participants who provided suspicious responses—particularly in open-ended fields—they were able to recover acceptable psychometric properties of the well-established Big Five personality inventory and replicate expected findings. This highlights the value of response screening as a quality-control tool.

To address these data-quality concerns, scholars have also increasingly turned to platforms that explicitly prioritize the recruitment of participants for human-subjects research, such as Prolific and CloudResearch (Peer et al., 2017). These platforms offer features aimed at improving data quality, such as prescreening filters and participant ratings. A recent evaluation by Douglas et al. (2023) suggested that these platforms have merit. By using a battery of quality-control measures—including instructed-response items (e.g., "select strongly agree"), prompts requiring participants to leave a field blank, detectors for implausibly fast responses, and suspicious internet protocol (IP) and geolocation flags—they found that participants were more likely to pass all quality checks on Prolific (68%) and CloudResearch (62%) than in a university subject pool (53%) or on MTurk (29%).

However, in November 2022, several months after Douglas and colleagues (2023) submitted their investigation for peer review (April 2022), OpenAI released a demo of ChatGPT. It quickly gained popularity along with a host of similar chatbots based on large language models (LLMs). These tools now pose a serious threat to the integrity of crowdsourced data. First, LLMs make it dramatically easier for fully automated bots to generate fluent, on-topic text that can evade open-ended attention checks. This raises the alarming possibility that entirely fabricated participants may go undetected, contaminating data sets with artificial intelligence (AI)-generated responses that appear indistinguishable from genuine human input.

Second, and nearly as troubling, is that real human participants can now use AI chatbots as on-demand assistants, cheating by outsourcing reasoning tasks, factual recall, or open-ended writing to a chatbot. With minimal effort, participants can generate articulate, accurate, and contextually appropriate responses that misrepresent their actual knowledge, understanding, or engagement with a task.

Whether AI-generated responses come from automated bots or human participants, they threaten the validity of crowdsourced science. During the 2018 wave of MTurk bot responses, researchers showed that inclusion of fraudulent responses could greatly increase the amount of noise introduced into survey data, decreasing the reliability of validated measures and weakening predicted relationships (Ahler et al., 2025; Chmielewski & Kucker, 2020). If fraudulent users are able to avoid detection by using LLMs to answer open-ended questions, this problem could resurge.

AI-assisted responding poses a validity threat when research questions concern unaided human cognition or behavior. It does so through two mechanisms. First, if an experimental manipulation involves a generative task—such as writing an essay or reflecting on personal experiences—but a participant outsources that task to an AI tool, they bypass the psychological processes that the manipulation is designed to engage. Second, if participants use AI to answer outcome measures, such as posttests designed to assess learning, these responses will likely be unaffected by experimental manipulations and may also outperform unaided responses. These processes should introduce noise into experiments, reducing observed effect sizes and increasing the sample sizes needed for researchers to test hypotheses. Without accurate detection and deterrence strategies, AI-assisted cheating has the potential to proliferate and diminish the internal validity of crowdsourced research.

As experimental psychologists who use the Prolific participant pool to study human learning and motivation, we recently began noticing that some of the open-ended responses in our studies began to take on characteristics of AI-generated text: verbose, grammatically correct

responses that could be strikingly similar to each other. To address this concern and better screen open-ended responses in our crowdsourced research, we built a lightweight and easily implementable JavaScript tool for Qualtrics that tracks the alignment between participants' keystrokes and their responses to open-ended questions.

The Present Research

Here, we report the results of three studies in which we used this tool, conducted in May, June, and July 2025. In the present research, we had four goals. The first was to deploy and evaluate a keystroke-based AI-detection tool. We tested whether a JavaScript keystroke-tracking tool could detect AI-assisted or outsourced responding on assessments in crowdsourced research. Because the studies included code-writing and statistical-reasoning tasks—skills that LLMs can perform accurately—we expected that flagged cases would show higher assessment scores, providing evidence of the tool's predictive validity. The second goal was to estimate the prevalence of outsourced responding in our studies. After validating the keystroke-tracking tool, we used it to examine how common outsourced responses were on coding and statistical-reasoning tasks in our modern Prolific samples. The third goal was to compare the tool with existing integrity checks and examine concurrent validity. We compared the tool with established measures of inattention and data integrity, including rapid completion times, suspicious IP addresses, and duplicate geolocations. This allowed us to assess concurrent validity (whether detected outsourcing correlates with other indicators of low-quality data) and determine whether keystroke logging identifies problematic participants who would otherwise go undetected by standard attention checks. The fourth goal was to explore potential effects of AI-assisted cheating on statistical power in crowdsourced studies. Using a simulated power analysis, we tested how the levels of cheating observed in our studies that were caught by keystroke tracking but not other standard quality checks might affect the sample sizes required to detect hypothesized effects in crowdsourced research.

Method

Studies 1 through 3 were all conducted on Prolific to test the impact of replacing lectures with practice problems and feedback. Here, we analyze data from these studies for a secondary purpose: to address questions about screening for AI use and preserving data integrity. Both the original research and this secondary analysis were approved by the Institutional Review Board at Carnegie Mellon University.

Table 1. Demographic Information, Studies 1 Through 3

Measure	Study 1	Study 2	Study 3
Age (years)	39.2	38.5	39.2
Gender			
Female	50%	52.5%	45.7%
Male	50%	46.8%	53.5%
Another identity	0%	0%	0%
Information not available	0%	0.7%	0.8%
Race/ethnicity			
Asian	3%	4.3%	5.5%
Black	20%	22.3%	18.9%
White	70%	64.5%	64.6%
Multiracial	4%	6.3%	6.3%
Another identity	2%	2.3%	2.4%
Information not available	1%	0.3%	2.4%

Participants

Participants were recruited via Prolific, screened to be at least 18 years old and residing in the United States. Data collection for each study continued until the preregistered number of participants had consented, completed the study, and met the inclusion criteria, which included checks for potential AI-assisted responding. Each study was designed to last approximately 30 min, and participants whose data met authenticity standards were paid $6.00. The target sample size was 300 for Studies 1 and 2 and 250 for Study 3. Before exclusions, 340 participants completed Study 1, 324 completed Study 2, and 266 completed Study 3. Before responses were screened for AI-assisted cheating, two participants were excluded from Study 1 for providing similar open-ended answers on the posttest at approximately the same time. Demographic information for retained participants is summarized in Table 1; demographic data were not available for excluded participants.

General procedure

All studies were administered via Qualtrics. In Study 1, participants were introduced to the basics of multiple regression, and Studies 2 and 3 provided an introduction to programming with Python. After providing informed consent, they completed a multiple-choice pretest assessing their knowledge of the relevant domain (statistics or computer science) and a questionnaire assessing their motivation and metacognition. Next, participants were randomly assigned to an instructional condition that included either a recorded lecture (Studies 1–3), practice problems with feedback (Studies 1–3), or worked examples (Studies 2 and 3). Following the instructional phase, participants completed a second motivation and metacognition questionnaire and an open-ended posttest assessing the material covered in the study.

Steps taken to limit use of generative AI

During the study, we took several steps to limit participants' use of generative AI. First, participants were informed on the consent form that the study contained embedded checks for AI use and that the research team would reject submissions in which AI use was detected. Participants saw a version of this message again on the first page of instructions for the study, which stated,

> **When you are asked questions, please don't use any outside tools like ChatGPT or Google.** This study has several built-in detectors to identify responses that use A.I., and we will reject tasks where participants are using A.I. tools like ChatGPT.
>
> **Your compensation for the study is not related to whether you get questions correct**, and we want to understand what you know. Use of tools like ChatGPT ruins our study.

After reading this statement, participants were required to confirm their understanding by typing "I will not use tools like Google or ChatGPT for any part of this study" in a text box at the bottom of the instructions page.

In addition, to limit cheating, we used JavaScript to disable participants' ability to copy test questions from the study, making it more difficult for them to look up answers. We did not disable pasting of text because we worried that if we did so, participants would be more likely to cheat by paraphrasing LLM-generated text, a pattern that would be more challenging to detect than pasting.

Measures

Although each study included a broader set of measures (described in their respective preregistrations), in this article, we focus on the measures most relevant to evaluating the implications of AI-assisted responding. Specifically, we attend to posttest performance (the primary outcome of each study), maintained situational interest (an established, Likert-style measure that allows us to examine if AI responding also predicts questionnaire validity), our new measure of AI-assisted responding on the posttest, and several established indicators of data quality for online studies.

Posttest performance. In each study, the posttest consisted of open-ended questions assessing the concepts covered in the instructional phase. Responses were scored by members of the research team using agreed-on rubrics. Study 1 focused on multiple regression, and Studies 2 and 3 assessed Python programming. For example, on the Study 1 posttest, participants were asked to identify the intercept of a line and interpret its meaning. On the posttest for Studies 2 and 3, participants wrote code involving "if" statements and variable assignment.

Maintained situational interest. Maintained situational interest in regression (Study 1) or programming (Studies 2 and 3) was measured on the postquestionnaire with four Likert-style items with the anchors *not at all* to *very much* (e.g., "How much would you like to learn more about programming?"). These items were adapted from Asher and Harackiewicz (2025).

Detection of AI-assisted responding. In the pre-LLM era, researchers typically identified AI-generated text by examining participants' open-ended answers directly and flagging those that were clearly nonsensical, irrelevant, or suspiciously brief (e.g., one-word responses; Douglas et al., 2023). LLMs have complicated this approach. Because LLM-generated text can often be fluent, coherent, and contextually appropriate, it can be indistinguishable from genuine human responses when judged on content alone. To address this issue, we adopted a different strategy, focusing not only on the content of responses but also the manner in which they were produced.

Keystroke-logging tool. To support this approach, we developed and embedded a lightweight JavaScript tool in Qualtrics to unobtrusively record participants' typing as they answered open-ended questions. We embedded this tool into the posttest of each study and into an open-ended practice session in Study 3. The tool logged keystrokes, paste actions, and "print screen" attempts for each response. We share code for the version of this tool used in this article and instructions for its use in Qualtrics on OSF: https://osf.io/f9w8c/files. An updated version of the tool (and instructions for setting it up) will be maintained on GitHub at https://github.com/the-oak-lab/keystroke-tracker.

After logging these actions, we compared logs with participants' final responses to assess whether an LLM or other external resource may have been used. Responses were flagged for potential outsourced assistance if the number of keystrokes was lower than the number of characters in the final text or if text was pasted into an answer box. All flagged cases were reviewed by a member of the research team, who examined both the keystroke record and the response content to determine whether to classify the case as involving outsourced responding. Flags were overturned, for example, if a participant appeared to paste and edit one of their prior responses or if the large majority of a response aligned with a participant's keystroke log and unlogged text was only a small proportion.

Validation of the keystroke-logging tool. To test the accuracy of the JavaScript keystroke logger, we conducted a preregistered validation study (see https://osf.io/u7kpt)

in which four research assistants (RAs) completed questions under controlled conditions. Each RA answered one question by typing a provided response and another by copying and pasting text from an LLM, simulating authentic and outsourced behavior, respectively. We screen-recorded all sessions and manually coded keystrokes for comparison with the tool's output. Results showed 99% character-level agreement between observed and detected keystrokes; one disagreement resulted from the tool recording a trailing "v" in addition to a "paste" action when an RA used "command + v" as a keyboard shortcut. All other disagreements were transpositions (e.g., "winter" logged as "witner"). All paste actions were correctly detected, and our classification criteria (paste action + anomalously low keystroke count) achieved perfect sensitivity and specificity in distinguishing outsourced from authentic responses. For full details of the validation study, see the Supplemental Material available online.

Indicators of general data integrity. We also examined whether flagged responses were associated with other indicators of low-integrity data that are commonly used in online research.

Time-based checks. Time-based checks were implemented by having a research team member complete each study as quickly as possible while still reading all questionnaire items, establishing a "fast-but-valid" benchmark (Douglas et al., 2023). Participants who completed surveys more quickly than this benchmark were flagged for potential low-quality responding. The thresholds were 40 s for the questionnaire items in Study 1 and 43 s for Studies 2 and 3, each covering 21 items. Total completion time was also recorded for all participants.

Questionnaire-based check. To detect inattentive responding on questionnaires, we conducted a long-string analysis on the study's 21 Likert-style items (Curran, 2016). For each participant, we counted their longest sequence of identical responses. The assumption behind this method is that inattentive participants may choose to complete questionnaires by selecting the same response option (e.g., strongly agree) for each Likert-style question. In prior research, participants exhibiting long strings have been found to complete surveys more rapidly, show lower even-odd consistency, and respond incorrectly to more attention-check items compared with attentive participants (Meade & Craig, 2012; see also Huang et al., 2012; Jones et al., 2023). Because there is no well-established accepted cutoff for what constitutes a suspiciously long string, we treated the measure as a continuous indicator of potential inattention.

Text-based check. To detect low-quality text responses, we adopted Chmielewski and Kucker's (2020) approach for identifying "unusual" responses. We flagged responses that consisted of single words that did not align with the question (e.g., "0" in response to a question asking a student to interpret a slope), nonsense phrases, or pure restatements of the question prompt. This indicator was dichotomous, marking whether participants provided at least one such response across their open-ended answers.

Metadata-based checks. We conducted two metadata-based integrity checks, following Douglas et al. (2023). First, we flagged participants with IP addresses that appeared more than once in a study. We also flagged participants with duplicate geolocations, in which multiple participants shared identical latitude and longitude coordinates. Although these duplicates can occur for legitimate reasons (e.g., college-student participants who complete a study from a shared campus wifi network), they may also signal fraudulent activity. This includes a single user operating multiple accounts or using a VPN to misrepresent the user's location and bypass a study's eligibility requirements.

Analysis

We analyzed the data from each study separately to see how the rates and consequences of AI-assisted responding varied over time and across statistical-reasoning (Study 1) and code-writing tasks (Studies 2 and 3). For each study, we used linear regression to test whether detected AI use was associated with continuous outcomes, including posttest scores, the length of participants' longest strings of identical questionnaire responses, and total questionnaire-response time. Logistic regression was used to examine associations between detected outsourcing and dichotomous indicators of low-quality data, such as being flagged for rapid responding, unusual text responses, or sharing an IP address or geolocation with another participant. When one or both groups had zero cases of a dichotomous outcome, precluding reliable logistic regression estimation, we used Fisher's exact test instead.

To calculate standardized effect sizes for cheating behavior on performance, we report Cohen's *d* (pooling the standard deviations for cheaters and noncheaters) when variances for these two groups are approximately equal. In studies in which variances are unequal, we report Glass's delta for a standardized effect size, dividing raw mean differences by the standard deviation of the noncheating group (Glass, 1976).

To test whether AI use on the posttest was associated with less reliable questionnaire data, we followed an approach taken by Chmielewski and Kucker (2020). Specifically, we tested whether the internal consistency of the situational interest scale (as measured by Cronbach's alpha) was significantly lower among participants who submitted AI-assisted posttest responses. Internal

Table 2. The Number of Participants in Studies 1 Through 3 Flagged for Possible Cheating by Each Detector and Exclusion Rates for Participants in Each Category

Category (flagged for)	*N* detected	*N* excluded	Rejection rate
Too few keystrokes typed only	30	7	23%
Pasted text only	37	2	5%
Both too few keystrokes and pasted text	83	71	86%

consistency differences were assessed using Feldt et al.'s (1987) method for comparing Cronbach's alpha, as implemented in the *cocron* package in R (Diedenhofen & Jochen, 2016). We also compared the alphas in each group with an established benchmark from prior research that used the same measure in a noncrowdsourced sample (Asher et al., 2025, Study 2, $\alpha = .92$ with $N = 338$ undergraduates).

Finally, to assess the potential impact of the observed levels of AI use on crowdsourced experimental research, we conducted a power analysis with simulated data, examining how the observed levels of AI use in Studies 1 through 3 might affect a researcher's statistical power to detect the effects of an experimental manipulation in a crowdsourced sample. The simulation modeled the sample size needed to maintain adequate (i.e., 80%) power when AI-assisted responses are undetected versus detected and removed.

Transparency and openness

In this article, we report how we determined our sample size, all data exclusions for each study, and measures related to AI detection. In each study's preregistration, we describe the full set of measures and all manipulations for each study. All data, analysis code, and research materials are available at https://osf.io/f9w8c. Data were analyzed using R (Version 4.5.0; R Core Team, 2025). Hypotheses and analyses about screening for AI use in Studies 1 through 3 were not preregistered.

Results

Prevalence of outsourcing: Keystroke data identified cheating for 9% of participants

The JavaScript tool automatically flagged participants for potential AI assistance based on two criteria: (a) evidence of pasting text into a response field or (b) a final response that was significantly longer than the number of detected keystrokes. Based on these two criteria, 67 participants in Study 1 (20%), 44 participants in Study 2 (14%), and 39 participants in Study 3 (15%) were flagged. Next, we manually reviewed each case to make a conservative final determination for each participant, rejecting their submission on Prolific only if evidence was conclusive that they outsourced a large component of a response to an outside resource.

As detailed in Table 2, the strength of evidence for cheating varied by flag type; the strongest evidence came when a participant was flagged for both pasting text and having missing keystrokes. We determined that participants were cheating in 85% of these cases. In contrast, flags for a single criterion were often inconclusive. Pasted text alone frequently appeared to be benign (e.g., participants pasting short code snippets to avoid retyping). Likewise, missing keystrokes without a paste action were difficult to conclusively attribute to cheating rather than potential keylogging errors or browser-based autocorrect or text-completion features.

After completing this manual review, we rejected a total of 80 participants (9% of the total sample) for cheating. This included 38 participants in Study 1 (10% rejected), 24 in Study 2 (8% rejected), and 18 in Study 3 (7% rejected).

Predictive validity: Flagged participants outperformed others by up to 1.5 SD

Figure 1 shows posttest performance in each study for the participants flagged for likely AI use versus participants who were not flagged.

As predicted, participants flagged for AI use consistently and substantially outperformed their peers, supporting the predictive validity of our keystroke-logging approach to AI detection. In Study 1, flagged participants scored 0.60 *SD* higher on the posttest than nonflagged participants, $t(336) = 3.38$, $p < .001$, with similar score variability in the two groups ($SD = 0.27$ vs. $SD = 0.30$).

In Studies 2 and 3, which both involved programming with Python, the effects of cheating appeared to be much stronger. Participants in Study 2 who were flagged performed homogeneously well on the posttest ($M = 97\%$, $SD = 9.5\%$), whereas participants who did not performed more than 1.5 *SD* more poorly, with much more variance in their responses ($M = 49\%$, $SD = 31\%$). Likewise, in Study 3, detected cheaters averaged 93% ($SD =$

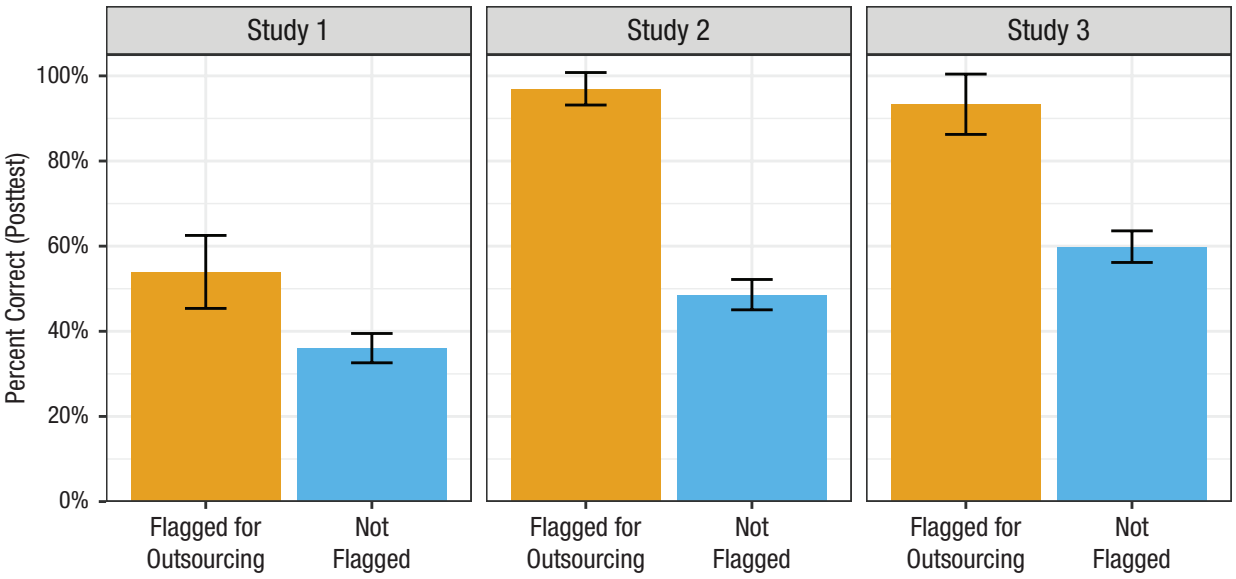

Fig. 1. Detected outsourcing and performance in Studies 1 through 3.

15%) compared with 60% (*SD* = 30%) for noncheaters, a difference of 1.12 *SD*, $t(264) = 5.13$, $p < .001$.

Concurrent validity: Outsourcing evaded standard checks but correlated with other quality indicators

To evaluate our keystroke-logging tool's unique contribution to maintaining data integrity in crowdsourced research, we compared its determinations against other common measures of data quality (see Table 3). The findings reveal that the keystroke analysis identified cheaters who would be missed by standard screening.

Critically, detected AI use was not consistently associated with three of the most common checks for inattentive responding: rapid completion times, long strings of identical answers, or short and irrelevant (i.e., "unusual") responses to the test questions themselves. This suggests that participants who outsourced their responses were not merely careless but were engaging in a different form of low-quality behavior that these specific checks failed to capture.

However, detected cheating was strongly correlated with other markers of problematic data, specifically, suspicious geolocations and unreliable questionnaire responses. Participants who showed conclusive evidence of outsourcing their responses were more than twice as likely to have IP addresses that were associated with duplicated geolocations. In Study 1, 85% of cheaters met this criterion (vs. 47% of noncheaters), $\chi^2(1) = 20.76$, $p < .001$, and in Study 2, 83% of cheaters did (vs. 28% of noncheaters), $\chi^2(1) = 29.69$, $p < .001$. In Study 3, the percentage of users with repeated geolocations dropped to 39% for cheaters (vs. 18% for noncheaters), $\chi^2(1) = 3.90$, $p = .048$. These numbers suggest that participants who outsourced responses may have been more likely to be using VPNs or proxy servers, perhaps to bypass geolocation filters and appear to be based in the United States.

It is notable that the overall proportion of users with duplicated geolocations fell between Study 1 (in which even 47% of noncheaters had a duplicated geolocation) and Study 3 (in which this number dropped to 18%). A logistic regression confirmed that Study 1 had significantly higher rates of duplicated geolocations than both Study 2, $\chi^2(1) = 25.05$, $p < .001$, and Study 3, $\chi^2(1) = 42.82$, $p < .001$. Study 2 also showed significantly higher rates than Study 3, $\chi^2(1) = 11.50$, $p < .001$, indicating a continued decline even between the two computer-science studies with nearly identical procedures. This suggests that the proportion of potentially fraudulent users on Prolific may have dropped between May and July 2025.

In addition to duplicated geolocations, detected outsourced cheating was also associated with reduced levels of internal consistency on questionnaire responses. When considering only the responses of participants who did not appear to cheat on the posttest, the measure of situational interest had a Cronbach's alpha of between .95 and .96 in all three studies. Among participants flagged for outsourced cheating, internal consistency was .11 units lower in Study 1, $\chi^2(1) = 17.40$, $p < .001$; .05 units lower in Study 2, $\chi^2(1) = 3.64$, $p = .056$; and

Table 3. Measures of Data Quality for Participants With and Without Detected Outsourcing in Studies 1 Through 3

Measure	Flagged for outsourcing	Not flagged	Test statistic	*p*
Study 1	(*N* = 38)	(*N* = 300)		
Responded too quickly	0%	1%	95% CI = [0.00, 4.05]	.605
Questionnaire time (minutes)	3.3 (*SD* = 2.6)	2.3 (*SD* = 2.1)	$t(336) = 2.69$	.007
Longest questionnaire string	4.9 (*SD* = 1.8)	4.4 (*SD* = 1.6)	$t(336) = 1.67$	.097
Short or irrelevant response	0%	5%	95% CI = [0.00, 2.38]	.382
Duplicated IP address	0%	1%	95% CI = [0.00, 42.42]	.999
Duplicated geolocation	84%	47%	$\chi^2(1) = 20.76$	.000
Alpha of interest scale	$\alpha = .84$	$\alpha = .95$	$\chi^2(1) = 17.40$	.000
Study 2	(*N* = 24)	(*N* = 300)		
Responded too quickly	4%	1%	$\chi^2(1) = 0.84$	.361
Questionnaire time (minutes)	2.5 (*SD* = 2.2)	2.2 (*SD* = 1.7)	$t(322) = 1.01$	.315
Longest questionnaire string	5.5 (*SD* = 1.9)	5.0 (*SD* = 1.9)	$t(322) = 1.39$	.165
Short or irrelevant response	0%	2%	95% CI = [0.00, 8.99]	.999
Duplicated IP address	0%	0%	—	—
Duplicated geolocation	83%	28%	$\chi^2(1) = 29.69$	.000
Alpha of interest scale	$\alpha = .90$	$\alpha = .95$	$\chi^2(1) = 3.64$	.056
Study 3	(*N* = 18)	(*N* = 248)		
Responded too quickly	5%	4%	$\chi^2(1) = 0.05$	.830
Questionnaire time (minutes)	2.7 (*SD* = 2.5)	1.9 (*SD* = 1.5)	$t(264) = 2.14$	.034
Longest questionnaire string	4.5 (*SD* = 1.5)	5.2 (*SD* = 1.8)	$t(264) = -1.21$	.227
Short or irrelevant response	0%	6%	95% CI = [0.00, 3.98]	.609
Duplicated IP address	0%	0%	—	—
Duplicated geolocation	39%	18%	$\chi^2(1) = 3.90$	.048
Alpha of interest scale	$\alpha = .76$	$\alpha = .96$	$\chi^2(1) = 24.16$	.000

Note: Dashes indicate cells in which statistical tests were not conducted because of zero variance. χ^2 statistics result from likelihood ratio tests. The 95% CIs result from Fisher's exact tests. CI = confidence interval; IP = internet protocol.

.10 units lower in Study 3, $\chi^2(1) = 11.35$, $p < .001$. Compared with the benchmark from prior research ($\alpha = .92$), detected cheaters showed significantly lower reliability in Study 1 ($\alpha = .84$, decrease of .08), $\chi^2(1) = 5.66$, $p = .017$, and Study 3 ($\alpha = .76$, decrease of .16), $\chi^2(1) = 7.83$, $p = .005$. In contrast, participants not flagged for cheating showed significantly higher reliability (by .03–.04 points, $p < .002$).

Overall, the keystroke-based measure demonstrated concurrent validity with other quality checks (75% of flagged AI users were also flagged by conventional methods, primarily repeated geolocations) but offered superior precision—flagging 9% of the sample versus 39%—and it provided direct behavioral evidence of outsourcing rather than indirect indicators of suspicion. For detailed analyses of overlap between detection methods, see the Supplemental Material.

Power analysis: Undetected outsourcing could increase Type 2 error rates by 50%

To assess the potential impact of the observed levels of AI use on crowdsourced experimental research, we conducted a simulated power analysis for an experiment including participants who used AI to cheat on a posttest either 0% or 10% of the time. We assumed that this experimental manipulation would produce an effect size of 0.4 *SD* (i.e., Cohen's $d = 0.4$) on the posttest among noncheaters and that cheaters would be unaffected by the manipulation, receiving high posttest scores regardless of experimental condition. We expect that cheaters should decrease statistical power through two mechanisms. First, if they do not respond to a manipulation, they will decrease its observed effect size. Second, if

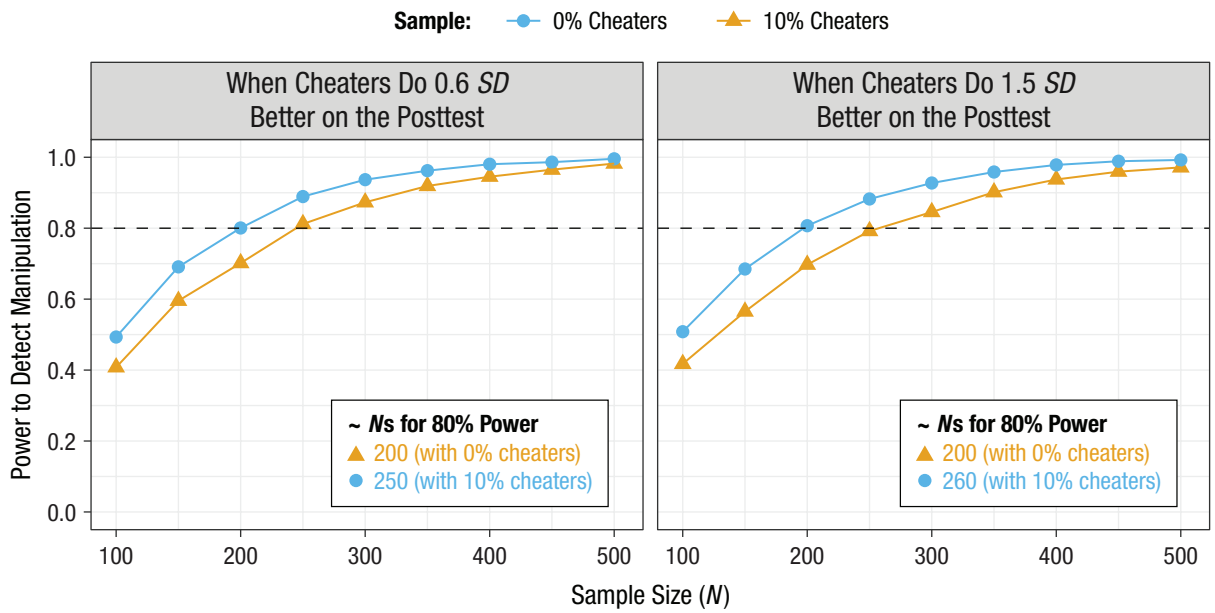

Fig. 2. Power analysis: sample size versus statistical power with and without artificial intelligence-assisted cheaters.

they perform substantially better than noncheaters, they could increase the variance of posttest scores, thereby introducing noise into the data that decreases the precision of estimates.

Because the average performance benefit of cheating varied in Studies 1 through 3 from approximately 0.6 *SD* (on a statistics posttest in Study 1) to 1.5 *SD* (on a programming posttest in Study 2), we simulated studies at both of these extremes. For the second study, we also narrowed the variance of the cheating group to one-third of noncheaters, reflecting what we observed in Study 2. The results of this power analysis are summarized in Figure 2.

Across all simulations, the introduction of 10% cheaters decreased the observed effect size of the experimental manipulation by a corresponding 10%, from $d = 0.40$ to $d = 0.36$, on average. This decrease in effect size and the additional variance introduced by cheaters had a moderate effect on statistical power. Assuming the smaller (0.6 *SD*) posttest impact of cheating, the necessary sample size for 80% power increased from $N = 200$ to $N = 250$, a 25% increase. With the larger posttest impact of cheating (1.5 *SD*), approximately 260 participants were needed for this same level of statistical power, a 30% increase in sample size. Critically, undetected outsourcing substantially increased researchers' risk of false negatives. A sample of 200 participants provided 80% power to detect the manipulation in a cheater-free environment but only 70% power when 10% of participants outsourced their responses. This represents a 50% increase in the Type 2 error rate (from 20% to 30%), meaning three out of 10 studies would miss real effects compared with two out of 10 in a cheater-free (or successfully screened) environment.

Discussion

In this article, we provide evidence that as of July 2025, AI-assisted cheating poses a substantial threat to crowdsourced research in the behavioral sciences. Across three studies conducted on Prolific—a platform often regarded as a "gold standard" for recruiting high-quality participants for human-subjects research (Douglas et al., 2023)—we estimated that approximately 9% of respondents used AI to answer open-ended statistics and coding test questions. We believe this figure is conservative: Because Prolific submissions flagged for AI use were rejected (preventing payment and lowering participants' approval ratings), we flagged cheating for only unambiguous cases, meaning less obvious cases may have gone undetected. It is also notable that this rate occurred despite active deterrence measures: disabling text copying, repeatedly stating that the study included AI detectors, and requiring written confirmation from participants that they would not use tools like ChatGPT.

Evidence for validity of the keystroke-logging tool

Our findings provide multiple forms of evidence about the validity of the keystroke-logging approach. The validation study provided evidence that the tool functions

as intended, demonstrating 99% character-level agreement with observed behavior and perfect classification accuracy under controlled conditions. Studies 1 through 3 provided evidence of predictive validity: Flagged participants substantially outperformed nonflagged participants (by 0.6 *SD* to 1.5 *SD*), consistent with the superior performance of LLMs on coding and statistical-reasoning tasks. We also found evidence of concurrent validity: Outsourcing was associated with other markers of problematic data, including duplicated geolocations and reduced questionnaire reliability. These patterns suggest that generative AI may be enabling generally inattentive or fraudulent users to complete crowdsourced studies undetected.

Importantly, the tool identified problematic participants who were not consistently flagged by standard attention checks, such as rapid completion times, low-quality text responses, or long strings of identical questionnaire responses, demonstrating its unique contribution to data-quality screening. The tool's core capability—allowing researchers to detect copy-paste behavior and mismatches between keystrokes and final responses—makes it broadly useful for identifying any form of outsourced responding, including unauthorized use of online resources in addition to AI use.

Implications for statistical power and data quality

A power analysis suggests that if unaddressed, the levels of AI-assisted cheating we observed could reduce effect sizes in experimental research by roughly 10%, requiring sample sizes up to 30% larger to maintain adequate statistical power. Without removing these cheaters, studies would be much more likely to miss real effects when they exist. In our simulations, the risk of such false negatives increased by about half.

However, it is important to recognize that AI-assisted cheating does not always simply add noise or weaken effects. If AI use is systematically related to experimental conditions—for example, if participants are more likely to cheat in a more cognitively demanding or less engaging condition—it could produce spurious findings rather than null results. Such patterns could artificially inflate effect sizes, reverse the direction of effects, or create misleading interactions with participant characteristics. Likewise, although we found that flagged participants provided less reliable questionnaire responses, inaccurate measurement does not always manifest in "worse" psychometric properties; it can sometimes produce deceptively high internal consistency when responses cluster together. AI-assisted responding threatens data validity in multiple ways, and its consequences depend critically on when, where, and why participants choose to use it. By tracking participants' keyboard inputs and comparing them with their submitted responses, these tools provide concrete, verifiable evidence of AI use—evidence that is difficult to establish through manual evaluation of LLM-generated text alone. If researchers can improve their detection and reporting of AI-assisted cheating, they will improve not only the integrity of their data but also the overall quality of research with crowdsourced samples.

Limitations

Although our findings point to clear risks and potential mitigation strategies, they should be interpreted within the context of several limitations. First, our keystroke-logging method detects copy-paste behavior and anomalously low keystroke counts, which are consistent with AI-assisted cheating but will also flag other forms of outsourced responding. Although participants could theoretically have copied text from websites or other sources, this seems unlikely: Our open-ended questions required specific responses to novel problems with customized variable names that would be difficult to locate online. LLM-assisted cheating is the most parsimonious explanation for the observed patterns.

Second, all three studies were conducted on Prolific between May and July 2025. Patterns of AI use may differ across platforms, participant populations, and time, particularly because both generative-AI tools and crowdsourcing-platform policies change. In fact, we observed a significant decline in duplicated geolocations over the course of our studies—from 51% in Study 1 to 32% in Study 2 and 20% in Study 3—which suggests that fraudulent participation on Prolific may have decreased during this period. Although alternative explanations (e.g., seasonal changes in participant-pool composition or even improved location spoofing by participants) cannot be ruled out, this trend could reflect improved detection, reporting, or moderation on the platform, underscoring the potential impact of ongoing vigilance and the value of reporting low-quality participants with platforms to strengthen participant pools.

Third, our experimental paradigm in all three studies, a learning session followed by a challenging posttest, may have influenced participants' motivation to use AI. Although we assured participants that compensation was independent of performance, the evaluative nature of a test and difficulty of coding and statistical-reasoning tasks could have caused some otherwise conscientious research participants to cheat. Our estimate of 9% outsourcing therefore reflects this specific research context rather than a universal base rate for Prolific. The prevalence of outsourcing likely varies considerably depending on task demands, study design, and other methodological

features. Studies using only Likert-style questionnaires or nonevaluative open-ended prompts, for instance, may show much lower rates of AI use, whereas studies with stronger performance incentives or more difficult assessments may show higher rates.

Finally, our detection method relied on keystroke logging. Although this approach provided concrete evidence of AI assistance in many cases, it likely missed some instances of partial AI use (e.g., hybrid responses, paraphrased AI content) and likely produced false negatives in which cheating was more subtle. It is also theoretically possible that sophisticated users could employ automated scripts to simulate human typing, thereby producing keystroke patterns that appear authentic. To detect this potential type of fraudulent responding, future work could explore solutions such as monitoring the cadence of keystrokes to better distinguish human versus AI responses.

Conclusion

Generative AI poses a new threat to the validity of crowdsourced behavioral research. Our three Prolific studies, conducted between May and July 2025, provide one of the first empirical estimates of AI-assisted cheating in this context, suggesting that at least 9% of participants outsourced their responses to open-ended test items despite active deterrence measures. This level of undetected cheating use can distort effect sizes and inflate sample-size requirements, but our findings also show that keystroke logging offers a practical way to detect it. Sharing such detection methods and reporting suspected cases to platforms may help strengthen participant pools and limit fraudulent activity.

As AI capabilities and platform policies evolve, the integrity of crowdsourced research will depend on continued vigilance, methodological adaptation, and communication between researchers and platforms.

Transparency

Action Editor: Katie Corker
Editor: David A. Sbarra
Author Contributions

Michael W. Asher: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

Gillian Gold: Data curation; Investigation; Methodology; Resources; Software; Writing – review & editing.

Eason Chen: Software; Writing – review & editing.

Paulo F. Carvalho: Conceptualization; Funding acquisition; Methodology; Writing – review & editing.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported in part by National Science Foundation Grants 1824257 and 2301130 to P. F. Carvalho and a Google Academic Research Award to P. F. Carvalho.

Open Practices

This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Michael W. Asher https://orcid.org/0000-0002-1006-8813
Gillian Gold https://orcid.org/0009-0007-9220-8915
Paulo F. Carvalho https://orcid.org/0000-0002-0449-3733

Acknowledgments

All data and code for this study are openly available at https://osf.io/f9w8c. Hypotheses, data-collection procedures, and analyses for the validation study were preregistered at https://osf.io/u7kpt. Data collection procedures for Studies 1 through 3 were preregistered at https://osf.io/28jtr (Study 1), https://osf.io/jwdtz (Study 2), and https://osf.io/ftwns (Study 3). Hypotheses and analyses about screening for artificial-intelligence use in Studies 1 through 3 were not preregistered.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459261424723

References

Ahler, D. J., Roush, C. E., & Sood, G. (2025). The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. *Political Science Research and Methods*, *13*(1), 1–20. https://doi.org/10.1017/psrm.2021.57

Asher, M. W., & Harackiewicz, J. M. (2025). Using choice and utility value to promote interest: Stimulating situational interest in a lesson and fostering the development of interest in statistics. *Journal of Educational Psychology*, *117*(4), 647–662. https://doi.org/10.1037/edu0000921

Asher, M. W., Sana, F., Koedinger, K. R., & Carvalho, P. F. (2025). Practice with feedback versus lecture: Consequences for learning, efficiency, and motivation. *Journal of Applied Research in Memory and Cognition*, *14*(3), 355–368. https://doi.org/10.1037/mac0000205

Buhrmester, M. D., 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. https://doi.org/10.1177/1745691610393980

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. https://doi.org/10.1177/1745691617706516

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. https://doi.org/10.1177/1948550619875149

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

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. https://doi.org/10.2308/bria-18-044

Diedenhofen, B., & Jochen, M. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, *11*, 51–60.

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

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Jones, E. A., Wind, S. A., Tsai, C.-L., & Ge, Y. (2023). Comparing person-fit and traditional indices across careless response patterns in surveys. *Applied Psychological Measurement*, *47*(5–6), 365–385. https://doi.org/10.1177/01466216231194358

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2020). *Assessing the risks to online polls from bogus respondents*. Pew Research Center. https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2020/02/PM_02.18.20_dataquality_FULL.REPORT.pdf

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

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. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Price, M., Hidalgo, J. E., Kim, J. N., Legrand, A. C., Brier, Z. M. F., Van Stolk-Cooke, K., Hughes Lansing, A., & Contractor, A. A. (2024). The cyborg method: A method to identify fraudulent responses from crowdsourced data. *Computers in Human Behavior*, *157*, Article 108253. https://doi.org/10.1016/j.chb.2024.108253

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007

Tomczak, J., Gordon, A., Adams, J., Pickering, J. S., Hodges, N., & Evershed, J. K. (2023). What over 1,000,000 participants tell us about online research protocols. *Frontiers in Human Neuroscience*, *17*, Article 1228365. https://doi.org/10.3389/fnhum.2023.1228365

Webb, M. A., & Tangney, J. P. (2024). Too good to be true: Bots and bad data from Mechanical Turk. *Perspectives on Psychological Science*, *19*(6), 887–890. https://doi.org/10.1177/17456916221120027

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. https://doi.org/10.1177/1948550617703168

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. https://doi.org/10.1037/pspa0000056