Who Broke Amazon Mechanical Turk? An Analysis of Crowdsourcing Data Quality over Time

Catherine C. Marshall, Partha S.R. Goguladinne, Mudit Maheshwari, Apoorva Sathe, Frank M. Shipman
Department of Computer Science and Engineering
Texas A&M University
College Station, Texas USA
shipman@cse.tamu.edu

ABSTRACT

We present the results of a survey fielded in June of 2022 as a lens to examine recent data reliability issues on Amazon Mechanical Turk. We contrast bad data from this survey with bad data from the same survey fielded among US workers in October 2013, April 2018, and February 2019. Application of an established data cleaning scheme reveals that unusable data has risen from a little over 2% in 2013 to almost 90% in 2022. Through symptomatic diagnosis, we attribute the data reliability drop not to an increase in bad faith work, but rather to a continuum of English proficiency levels. A qualitative analysis of workers' responses to open-ended questions allows us to distinguish between low fluency workers, ultra-low fluency workers, satisficers, and bad faith workers. We go on to show the effects of the new low fluency work on Likert scale data and on the study's qualitative results. Attention checks are shown to be much less effective than they once were at identifying survey responses that should be discarded.

CCS CONCEPTS

• Information systems -- World Wide Web -- Crowdsourcing

KEYWORDS

Crowdsourcing; data quality; data cleaning; Mechanical Turk

ACM Reference format:

Catherine C. Marshall, Partha Goguladinne, Mudit Maheshwari, Apoorva Sathe, Frank Shipman. 2023. Who Broke Amazon Mechanical Turk? An Analysis of Crowdsourcing Data Quality over Time. In *Proceedings of ACM Web Science (WebSci'23). ACM, New York, NY, USA, 10 pages.* https://doi.org/10.1145/1234567890

1 Introduction

For over fifteen years researchers have used Amazon's crowdsourcing platform, Mechanical Turk (referred to as AMT or MTurk), to collect different types of participant data. While there were initially concerns over data quality, Human Intelligence Task (HIT) design strategies were developed to identify and remove bad-faith work [25]. For surveys, the most popular design remedies have included attention checks, reading comprehension questions, and threshold times for good faith HIT completion [21]. Other types of HITs have relied on techniques like comparison with gold sets (known answers) and inter-rater agreement [2].

As time passed, workers became more aware of reliability checks [26]. As a result, they answered these questions carefully and held surveys for extra time to ensure that requesters approved their work. More recently, survey designers have adopted measures to vet workers based on IP addresses or device location [1][24].

Standard approaches to identifying bad faith submissions have led researchers to remove an ever-increasing number of responses. E.g., in 2018, Kaplan et al. reported removing about 10% of their data (40 out of 400 completed surveys) based on responses to open-ended questions [23]. In 2022 research that was directed at work quality, Dupuis et al. reported a HIT failure rate of 27% even with relatively stringent worker prequalification [12].

Table 1 illustrates this trend using a survey HIT that we have fielded four times over the past decade. The HITs maintained the same participation requirements (US-based workers with a 95% acceptance rate). The surveys fielded in 2018, 2019, and 2022 were identical to one another, adjusted in minor ways from a HIT we originally posted in 2013 [38]. Datasets were cleaned using a point system and reconciled using inter-rater agreement [27]. Submitted work that we needed to discard using this cleaning regimen jumped from 2.4% in October 2013 to 88.8% in June 2022.

Table 1. Recent acceleration of bad responses to the same survey fielded as a HIT on AMT

Survey Date	Total Responses	Bad Responses	Percentage of Bad Data
October 2013	250	6	2.4%
April 2018	500	26	5.2%
February 2019	500	62	12.4%
June 2022	250	222	88.8%

Unusable data can arise for a number of reasons, some under requesters' control (e.g. survey design [40]; task clarity [15]; payment [29]; and day-time-and-season strategies for fielding HITs to maximize worker diversity) and some outside of requesters' control (e.g. worker-side tools or bots [24]; platform-related limitations on worker screening [11]; and changes in the worker population over time [34] [10]).

WOODSTOCK'18, June, 2018, El Paso, Texas USA
 978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/1234567890

Because participant data collected using AMT has been valuable and surprisingly rich, many researchers (including us) are reluctant to stop using the platform. Instead, it seems worthwhile to diagnose what lies behind the current spate of bad data and develop mitigation strategies. In this paper, we analyze different types of problematic survey responses and describe how they have changed over time. We go on to assess the effectiveness of attention checks as a way of vetting work. We show evidence of what appears to be worker collaboration or workers controlling multiple identities on the AMT platform and examine possible effects of language proficiency shortfalls. Finally, we discuss future work and what it will take to revive the reliability of AMT and other related crowd platforms in which workers perform important microtasks for pay or comparable rewards.

2 Method

Our experiences with AMT in 2019—as well as an informal examination of other researchers' survey results—made us wonder whether something important had changed to decrease the reliability of US-based work. Colleagues gave us access to datasets from surveys they ran using US participants recruited on Dynata and AMT prior to pandemic workplace closures in 2020. We noticed some puzzling patterns in their data that matched what we had observed in our own data. Not only did the responses to open-ended questions exhibit known indicators of bad faith work (e.g., non-sequiturs such as "GOOD" and "NICE"); they also included more ungrammatical and misspelled statements, along with canned text that didn't answer the questions. Our colleagues had collected worker locations and IP addresses as a secondary assurance that workers were in the US. A few surveys had been completed on cloud computing platforms or using VPNs, but most originated from normal broadband services in the US. An apparent increase in bad data led us to believe that we might need to refine our data cleaning methods.

Subsequently, we designed and piloted four short surveys on different topics with the aim of identifying new genres of questions that would make AMT data easier to algorithmically clean. But the results again had more bad responses than we had anticipated. The cleanability issues we set out to investigate were subsumed by the dramatic overall decline in reliability.

Instead of trying to diagnose the questionable results from the new pilot surveys, we decided to compare contemporaneous AMT survey data to a historic baseline. So we re-ran a survey that we had developed as part of a well-documented larger study of the ownership and control of personal data [28]. The HIT and the method we used to develop it are described in [26] and [27]. The data collected in October 2013 provides us with a baseline. The 2018 and 2019 data serves as a window onto the uptick of unreliable work. Together, the previous surveys provided 1250 survey responses, both good and bad. Would there be a further increase of unusable data in 2022?

In addition to 25 Likert scale judgments, the survey asked workers six open-ended questions, mainly about personal content creation and removal on social media platforms like Facebook and LinkedIn. A final open-ended question asked about institutional archiving of social media by the Library of Congress; in the past that question had garnered strong opinions about whether a person's public self-presentation on Facebook should be regarded as ephemeral or part of the historical record. For more detail on previous results, see [27] and [38].

We have included the HIT instructions and questionnaire as supplementary material for this paper. Readers interested in the scenarios and open-ended questions are encouraged to read the HIT to better understand the analysis presented in this paper.

To gather contemporaneous data, we fielded this survey as a batch of 250 HITs on Monday, June 13, 2022 at 1pm CDT (6PM GMT). The survey was offered to US-based workers with a 95% HIT acceptance rate to duplicate what we did in 2013, 2018, and 2019. This time, the batch completed in about two hours. We used the points system described in [27] to flag bad data across the questions, including:

- erroneous responses to two reading comprehension questions
- individual unanswered questions
- nonsense answers to open-ended questions, and
- HITs completed faster than a 6.5 minute threshold.

In the past, the threshold of two allowable errors eliminated most of the questionable submissions, while not cherry-picking data for analysis. This cleaning regimen retained surveys that were completed very quickly but were otherwise fine. Sometimes fast completion signaled intense focus or a work style in which workers began the survey before they accepted the HIT. This points system also retained surveys with minimal answers to open-ended questions, but correct responses to the two attention checks.

Because some of the flags are judgment calls (e.g. what constitutes a nonsense answer), the completed surveys were assessed independently by two judges. The judges initially agreed on the disposition of 241/250 HITs. The nine HITs with differing judgments were resolved through discussion. This process resulted in 222 surveys that under normal circumstances we would have discarded. The 222 discarded surveys are the subject of our analysis in this paper. We go on to discuss the results of this analysis.

3 Results

We began our data analysis by examining the types of bad and hard to clean data we encountered in 2022 and discuss whether each is an ongoing problem or whether it is new. We then look at the effect of bad data on workers' Likert scale judgments. Because attention checks still form the backbone of many data cleaning strategies, we assess their current effectiveness. We also examine recurring misinterpretations of questions as part of our detective work into the possible sources of bad data provided in good faith.

3.1 Types of open-ended responses

Minimal responses. There have always been questions about how complete open-ended answers need to be to demonstrate that a worker is completing the survey in good faith: is it enough to answer the attention checks correctly, but respond to most or all the open-ended questions with "*N/A*" or "*I don't remember*"? This type of data generally represents the work of *satisficers*, workers who

provide minimal responses to open-ended questions. Satisficers may be taking many surveys in a work session and may visit multiple survey-for-pay sites in a day [18].

We encountered less satisficing in 2013 [26]. Although a few workers gave minimal answers to open-ended questions, they usually elaborated on at least one question. Only one worker's data was discarded because he didn't respond to open-ended questions. By the next time we fielded the survey in 2018, satisficing was more common (c.f., [14])

From a requestor's perspective, satisficing data may require that acceptance thresholds be revisited: How can we distinguish satisficing from incomplete or bad-faith responses? E.g., when a worker is asked about the last time they removed information from social media and why, and they respond "*a picture*," it is difficult to tell whether they understood the question's intent or whether they were just trying to save time. A more nuanced threshold can help discriminate between acceptable levels of satisficing (the survey-taker is working quickly but seems to be attending to the questions) and unacceptable levels of satisficing, where workers seem to assert that none of the questions apply to themselves.

By 2018, not only did satisficing increase, we also began to detect other types of suspicious answers to open-ended questions, including some that other researchers thought to be symptoms of bad faith work [8][6]. More significantly, we observed an increase in answers that seemed to point to a lack of English proficiency (we required survey-takers to be US-based English speakers).

Incomplete or bad faith responses. In the early days of AMT, most survey-takers were reluctant to submit bad faith work for fear that it would be rejected, thus making them ineligible for future work. Workers who found the survey too long, the pay too low, or the topic irrelevant to their interests would just return the HIT to the pool for another worker to do. Studies found the reputational threat sufficient to eliminate the need for attention checks. [32]

As time went on, occasionally answers to open-ended questions were conspicuously meaningless (e.g. "*fgfg*"). Workers also began to enter positive words or phrases (e.g. "*good*" or "*nice survey*"), which was likely a strategy learned from other workers. In 2018, of the 26 discarded surveys (5.2% of 500 returned), only six workers gave genuinely inscrutable answers to the open-ended questions. E.g., one survey taker answered one of the two attention checks (a reading comprehension question seeking a one-word answer), "*Using ingenuity and hard work along with his father's customer base and support, Harry built a fertilizer ammonization plant*," a non-sequitur so profound we assumed it was a joke or text from the worker's cut buffer. It wasn't until 2019 that we noticed significant examples of apparent bad faith work. As we examined the work more carefully, it seemed more likely that some of this work stemmed from workers who were not fluent in English.

Low and ultra-low fluency data. Low fluency data, responses that demonstrate a lack of ready familiarity with a survey's language, is a relatively new phenomenon on AMT. In our case, we have relied on Amazon to limit a HIT's availability to US-based workers fluent in English. We cross-checked this requirement within the survey, but in the 2019 and 2022 surveys, workers who were not fluent in English attempted to answer open-ended questions in a good-faith effort to have their work accepted.

Low fluency responses may include poorly constructed or ungrammatical language, or content copied from websites or from the survey itself. Naturally, fluency exists along a continuum: some workers can produce answers in their own words (possibly with the aid of a translation tool in one or both directions), while others must locate likely-looking text on the Web. We began to see answers that appeared to be the result of using a survey question as a search engine query and copying top results or snippets surfaced by the Search Engine Results Page (SERP) into the response.

We distinguish low fluency work (in which open-ended results are poorly stated, but unique) and ultra-low fluency work (in which open-ended results are plagiarized from the Web or from SERP snippets). Is low fluency data still usable if it is produced in good faith? Do low fluency workers provide usable results to other types of questions (e.g. Likert judgments)? It's not unusual for reading comprehension to exceed language production.

If workers meet domain familiarity requirements (in this case, are Facebook users), do they need to be eloquent about their opinions? To answer this question, we must revisit the study's goals. Our surveys explored digital content ownership and copyright. To interpret answers, participants need to be working within the same legal system as the researchers.

We have re-examined earlier data. Low fluency data was not evident in the 2013 responses. It was beginning to be an issue in 2018 and 2019. In 2022, the amount of low and ultra-low fluency data was alarming. We returned to multiple related datasets we collected between 2010 to 2013 that required US-based Turkers to answer open-ended questions and discovered that low fluency data was all but nonexistent in the early years of the MTurk platform.

The novelty of low fluency responses made them difficult to detect in 2018. E.g., one worker answered the question about archiving Facebook, "*The Internet Archive visual arts residency, which is organized by Amir Saber Esfahani and Andrew McClintock, is designed to connect artists with the archive's 40 petabytes of digitized materials. Over the course of the yearlong residency, visual artists create a body of work which culminates in an exhibition*." A response like this, cribbed from *Fast Company*, stood out at data coding time, but not when the raw data was marked for cleaning. Other low fluency answers might strike a researcher as careless writing. In 2018, a participant answered a question about unusual uses of Facebook: "*add the face book account in email address was enable then copy of some content in photo, videos saved the idem*." In isolation, the response just seems hastily composed. But in the context of the rest of this participant's open-ended responses, the worker's level of English proficiency is hard to miss.

Before they were common, low fluency responses might not be flagged as symptoms of a larger problem. As they became more common, low fluency responses may continue to go unnoticed because they seem within the realm of normal good faith answers.

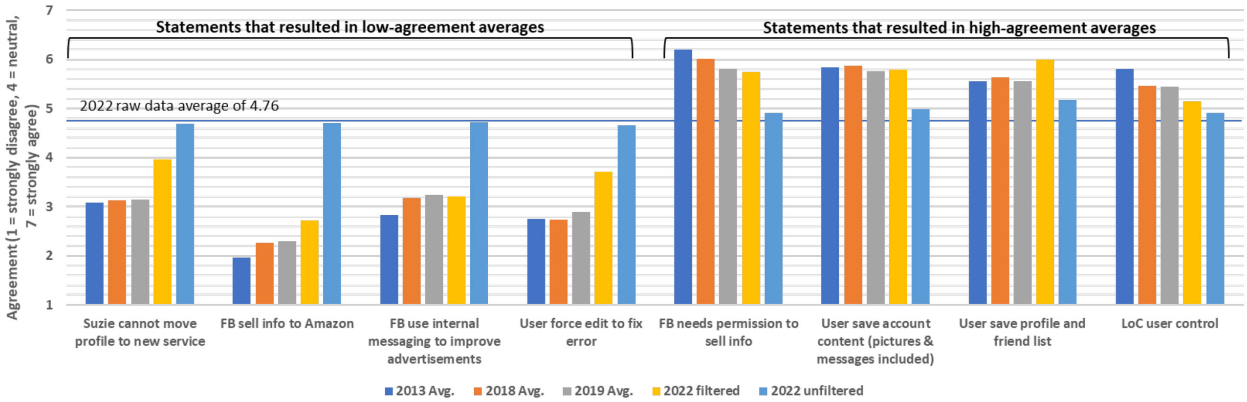

Figure 1. Trends from cleaned survey data from four survey fieldings (2013, 2018, 2019, 2022) are compared with unfiltered 2022 data (the light blue bar). Average values from eight Likert scale judgments are shown: four that elicited negative reactions in the past and four that elicited positive reactions.

3.2 Assessing the Effect on Likert Scale Data

How does the decrease in worker fluency affect responses to questions that aren't open-ended? We compared Likert judgments from 2013, 2018, and 2019 with those obtained in 2022, both filtered according to the prescribed cleaning regimen and unfiltered (i.e. all completed surveys were included).

Across the 25 7-point Likert scale questions, the average rating for the unfiltered data from 2022 varies from a low of 4.54 to a high of 5.18, resulting in a range of 0.65. In other words, the judgments are all mildly positive and don't vary much. From past experience with this survey, we anticipated the per-statement Likert judgment averages would vary far more. In fact, the range of per-statement averages was 4.23 in 2013, 3.74 in 2018, and 3.51 in 2019. Looking only at the data left after cleaning, the range for 2022 was still only 3.04. Thus, the post-cleaning data shows a clear downward trend. Our initial interpretation is not that participants are more apathetic about content they contributed to social media platforms, but that engagement with the survey itself is lower. Thus reduced English fluency—coupled with other survey participation problems—seems to influence the Likert judgments, too.

Taking a more detailed look at this problem, Figure 1 uses eight example Likert judgments about social media content to illustrate the problem with the 2022 survey's quantitative data. The right side of the figure shows four Likert scale judgments that elicited more positive average responses in the past (e.g., Facebook should get user consent before it sells personal information [27]), while the left side shows four statements where the average response used to disagree with the statement (e.g., a LinkedIn user should be able to force another user to fix errors in their profile if it causes the first user competitive harm [38]).

Figure 1 shows that from 2013 to 2019, only small shifts in judgment were detected in these eight example cases. The 2022 filtered data seems poorly aligned with the emerging patterns and trends (although the post-cleaning dataset is much smaller than our previous datasets). The 2022 data deviates substantively from prior values and trends and suggests that it would possible to misinterpret new survey results even after a rigorous cleaning process.

Overall, these results indicate that quantitative feedback of the type provided by Likert-scale judgments and similar sorts of multiple-choice questions are likely to be influenced by the fluency changes we observed in the analysis of the open-ended questions.

3.3 Do Attention Checks Still Work?

Attention checks became a standard element of crowdwork after Kittur et al. demonstrated their effectiveness in 2008 [25]. Not long afterward, worker forums began discussing them, pointing out those that were particularly tricky or clever. More importantly, workers alerted other workers of their presence [26][20]. This means that attention checks may be answered correctly even if other parts of the survey are unreliable.

Recently, researchers have suggested that attention checks should be redesigned as casual open-ended questions (e.g., "*What did you have for breakfast today*?") An open-ended attention check of this sort can tell us something about the workers, in addition to ensuring that they are reading carefully. As political scientist Tim Ryan pointed out in a blog post from 2020 [35], open-ended questions that ask about everyday topics may reveal workers' cultural characteristics. We saw this effect in our 2022 data: some workers used a Britishism, *fresher,* to refer to a college freshman. However, even open-ended attention checks may provide ambiguous results.

Table 2. Correct answers to attention checks relative to data quality for the rest of the survey

Attention check	correct responses for discarded data (out of 222 discarded)	correct responses for good faith data (out of 28 kept)
q1	110 (50%)	27 (96%)
q2	82 (37%)	26 (93%)
both	69 (27%)	26 (93%)

In 2022, all but two of the workers taking our survey in good faith answered both reading comprehension questions correctly. But so did quite a few of the more suspicious efforts. Table 2 demonstrates that we would have kept bad data (and discarded good data) if we relied on one or both of the reading comprehension attention checks that appeared to function as designed back in 2013.

Even if we required workers to answer both attention questions correctly, we would have retained more than triple the amount of 2022 data than we did using our point system.

The unusual number of questionable open-ended responses, coupled with the growing ineffectiveness of attention checks, led us to ask what happened to the reliability of the 2022 data.

3.4 Effect on Open-Ended Responses

When low fluency work first began to appear in 2018 and 2019, it was still sufficiently infrequent that we did not differentiate it from other good faith efforts unless we noticed plagiarism. But more recently, language fluency symptoms have become common enough to stand out as a specific phenomenon. Questionable responses may stem from a variety of language-related mishaps: misinterpreting HIT questions, using HIT questions as web queries and copying text from a web page or from a SERP snippet, and misinterpreting a question's genre (e.g. mistaking a question about personal practice as a general factual question).

Misinterpreting questions. Common idioms sometimes creep into survey questions to make them seem more approachable. Here's an example of an open-ended question that provoked a few mystifying answers in 2019: "*Describe the last time you removed content from a social network like Facebook or LinkedIn (especially if you removed more than one item).*" The phrase "*Describe the last time…*" is idiomatic, albeit common. In 2019, only 14/500 (<3%) of respondents answered the question using a literal time of day or date: "*This week*" or "*Maybe 5 years ago?*" or "*yesterday*." Only one of these 14 workers didn't seem to be working in good faith. At the time, these responses were puzzling, but seemed like they represented satisficing workers' efforts to avoid a complicated question. By 2022, this misinterpretation was more prominent. Answers like "*11:00 PM*" appeared in almost 10% of the responses.

Text from the web. Answers copied from the web were uncommon in past years. It's easy to see why: even if a worker was satisficing, it's more effort for a native speaker to locate a topically relevant web page and copy text from it than it is to type a brief answer reflecting the worker's own experience. If a worker chooses good candidate text, it can be difficult to detect this approach when plagiarism is rare. E.g., in the 2019 survey, one worker's response to the removal question we described above was: "*train journeys and lunchtimes were spent hopping from one app to another, cruising for attention in the form of likes. I'd open Facebook, then Instagram, then Messenger, and in the time it had taken me to look at the latter two there was a chance that something might have happened on Facebook*." On its face, the narrative seems sufficiently personal to pass muster. Upon further investigation, we found that the response was plagiarized from a *Guardian* article dated 14 March 2017.

Text from a SERP snippet. We distinguish a second type of plagiarism in which responses were probably copied from the SERP snippet that appears when a worker uses the question's text as a web query. These responses seem different than those which require navigation because they may signal even lower fluency.

If workers were using a tool to support this type of plagiarism, we might expect workers' answers to overlap. Although there was occasional duplication of responses, there was also sufficient variety to suggest the process was not fully automated by any of the common worker-side tools; localization, personalization, and search engine differences may also help account for variety, although when the authors Googled the entire question, it was easy to see a potential path from question to response.

Few workers restated plagiarized answers in personal terms. E.g., when participants were asked, "*What types of things do you use the Internet for? What do you spend the most time doing on the Internet?*", five workers (out of 250) responded, "*92% of Internet users have used [the] Internet for sending e-mails and for using search engines. 83% for getting more information related to health or hobbies. 82% for searching directions. 81% for getting weather information*." Workers (or a bot) with greater English proficiency might have personalized or excerpted this as "I *have used [the] Internet for sending e-mails and for using search engines*" or "*for sending e-mails and for using search engines*" to make the plagiarism harder to detect. In practice, workers did little to make SERP snippets look like their own answers.

Out-of-genre answers. At the root of the most inexplicable answers were workers who mistook one genre of question for another. E.g., a reading comprehension attention check (in which the answer is in the HIT) can be mistaken for an attention check seeking a fact. In 2022, over 12% of the incorrect answers to the survey's second reading comprehension question were related to the popular US TV show, *The Simpsons* (see Table 3). Was this a cynical complaint about the second attention check? It seemed at the outset that it must be. Returning to the 2013 and 2018 raw data, there were no references to the Simpsons. In 2019, only 4 workers out of 500 referred to the show, a small enough number to flag the error without wondering why a worker referred to a popular TV show.

In 2022, at least 17 workers had used the attention question, "*What's the name of the company where Greg and Homer work?*" as a web search, misidentifying the attention check as a factual question. Because Google's SERP for the question features the snippet "*Apu misses his job, so he and Homer travel to India to persuade the head of the Kwik-E-Mart corporation to rehire him,*" extracted from a Wikipedia page, *Homer and Apu*, eight workers answered with the proper noun most proximate to corporation—Kwik-E-Mart. Others, mistaking the word "*company*" to mean human company or friends, arrived at the answer, *Homer and Apu*. But the Wikipedia page in question does not mention Montgomery Burns (three workers found this answer), nor do other English-language search engines we tried (e.g. Bing and DuckDuckGo) produce same SERP as Google. It seems that workers went to some lengths to come up with these incorrect answers.

Table 3. A categorization of bad answers to second attention check, a reading comprehension question "*What's the name of the company where Greg and Homer work?*" Parenthetical numbers refer to the number of times the answer occurred

Bad q2 answers	#	Examples
Companies appearing in the HIT	19	i.e. Facebook (8), Amazon (6), and LinkedIn (5)
Companies or employment sectors not appearing in the HIT	22	e.g. Health/HALTH/HELTH Care (6), HP Inc (4), IT/Infotech (4)
Answers from SERP and Wikipedia related to a Simpsons episode	17	e.g. Kwik-E Mart and variants (8), Homer starts to work for a friendly (4), Montgomery Burns (3)
Names appearing in the HIT	10	i.e. Greg/Gerg/Greg and co. (7), Homer and variations (3)
Names not from HIT	6	i.e. Gusray, Gurnsey, Bettis, Keira, jonshan, jhon
Text copied from HIT	31	e.g. Journalist (10), profiles at Xiblix (5), company (4), Facebook is her main communication hub. (1)
Text of unknown provenance	4	All the workers, cultural references, producing, office
Non-answers	31	e.g. No (2), Yes, N/A (6), not mentioned (1), left blank (17)

This attention check seemed almost too easy at HIT design time. The scenario and subsequent Likert scale statements mentioned the company name, Xiblix, four times. We expected workers might not remember the fictitious name, but they could readily scan prior text to find it. Even if low fluency workers had detected the first attention check and answered it correctly, many didn't identify the second one as a reading comprehension question whose answer would appear in the HIT. In some cases, they may have tried: the word *company* does not appear near the word *Xiblix*. Context should have helped, and in past years, it did.

3.5 Range of Compensating Strategies

Workers exhibited different strategies for coping with fluency deficits. Table 3 shows many of the common incorrect answers to the attention check question we just described. That some of the answers were so inappropriate may mean that workers were translating the question into a language they understood, then guessing at an answer.

Those who recognized the question as a reading comprehension check and understood English well enough to know what type of answer was expected apparently scanned the HIT for company names: Facebook (the only company referred to as such), LinkedIn, Xiblix (the correct answer), Circles, and Amazon are real and imagined company names in the HIT's brief vignettes. Proximity of these company names to the attention check seemed to make some wrong answers more likely than others. No one, e.g., chose a fictitious company, Circles, that appeared earlier in the HIT (one worker did so in each of 2013 and 2018, and two did in 2019).

Those who didn't understand that the question was a reading comprehension check but were somewhat fluent in English knew they needed to come up with a company name. Several of the companies that workers named are based in India (e.g. SSC Company Private Limited); others named a multinational (e.g. HP Inc.). Six workers answered with a variant of "Health Care" although there was no mention of a health care company in the HIT.

The high variability of answers makes it unlikely we can blame the workers' performance on a particular tool, bot, or technique. If so, they would've converged on a smaller, less varied set of wrong answers. Instead, they seem to have relied on different strategies for language deficits. Some may have used a translation tool; that would explain the alternate understanding of the word "company" to mean "the company one keeps" or friendship (which might lead them to answer "Greg" or other names). Others may have used an English language search engine—possibly understanding neither the question nor answer—to come up with likely sources of information. Still others may have made judicious use of the local search function to find likely text in a sentence containing both Greg and Homer (which is why 5 workers answered "*profiles at Xiblix*" rather than just "*Xiblix*").

The introduction of typos, spelling errors, and alternative capitalizations suggests that workers sometimes typed answers rather than copying them directly from search results.

4 Collaboration or identity fraud?

Overlapping unusual answers may signal that one person is completing multiple surveys, which in turn implies the worker controls multiple worker IDs, since workers are only allowed to check out one survey HIT from a given batch. Overlapping answers may alternatively point to worker collaboration (a phenomenon documented by Gray et al.'s fieldwork [16] and addressed by Checco et al.'s system [7]).

Table 4 provides an example of this phenomenon by pivoting on a single distinctive wrong answer to the first attention check, AC1, "*Susie is looking for a job as a _____?"* All five worker IDs replied with the word "*women*". This answer is both culturally inappropriate (we expected the profession named in the scenario, *journalist,* although the workers' answer is true by inference from the pronouns, name, and photo) and grammatically incorrect (the answer should be singular). Furthermore, the word "*women*" doesn't appear in the HIT. If this were a common misinterpretation of the attention question, surely it would have appeared in previous years, given raw data from 1250 previous workers.

Table 4 compares responses to four other questions, two open-ended questions (Q1 and Q2), the other attention check (AC2), and one Likert scale judgment that in the past elicited stronger disagreement than all other judgments. The answers are presented

Table 4. Example of a crowdwork cluster: signs of low fluency workers collaborating or controlling more than one account. The selection pivots around the answer "*Women*" in response to AC1. The questions in the column headings are an abbreviated form of those in the survey.

Q1: What types of things do you use the Internet for?	*AC1: Susie is looking for a job as a _____?*	*AC2: What's the name of the company where Greg and Homer work?*	*Q2: Would it be okay for an institution to archive the public contents of Facebook (including your stuff)? Why or why not?*	*Q3:Okay for FB to sell user data to Amazon (Likert judgment)*	*Time spent on HIT (in seconds)*	*Time HIT was accepted by the worker*
5G	WOMEN	GREG HOMERWORK	BEACAUSE IT IS A PROCESS OF FACEBOOK	7 - strongly agree	703 (~12 min)	13:30:50
5G,4G ARE OF INTERNET	WOMEN	GREG AND COMPANY	YES IT IS A PROCESS FOR THE COMPANY	7 - strongly agree	905 (~15 min)	13:34:16
4G,5G NETWORK IS A VERY FAST NETWORK	WOMEN	GERG AND CO	YES IT IS THE PROCESS OF THIS INSTITUTION	5 - slightly agree	352 (~6 min)	13:54:09
4G 5G ARE USED BY ME	WOMEN	GREG AND CO	YES IT IS A PROCESS OF THE INSTITUTION	6 - agree	277 (~5 min)	14:10:26
5g 4g are the networks are used by me	women	greg and co	yes it is process of the institution	6 - agree	2292 (~38 min)	14:07:42

in submission order; the surveys were completed the date the HIT was fielded (13 June 2022).

The coinciding answers involve significant misinterpretation of each question, but are less identical than they initially appear. The Likert judgments are all positive (contrary to our past findings, and consistent with our earlier observation about the positive-trending Likert judgments in 2022's bad data—see Figure 1).

Did the same person answer all five surveys under different identities? Or were three workers working together, filling out five surveys? The demographic characteristics differ. Two report they are women; three report they are men. Birth years are distributed across several decades as well.

The answers to the open-ended Q2 do not appear elsewhere in the 2022 survey responses (nor do they appear in other years) and suggest a different cultural norm than is common in the US. US workers' responses are ownership-centric, privacy-motivated, and often say they consider Facebook data to be ephemeral. The 2013, 2018, and 2019 responses tilt heavily toward "No", that it wouldn't be okay for an institution like the Library of Congress to archive public data from Facebook. Instead, these answers all say "yes" (although the first answer does not say yes explicitly). While the first two answers say that archiving is something Facebook already does ("*a process for the company*"), the final three seem to say that archiving is a function of the Library of Congress ("*a process of the institution*"). It's exceedingly rare for US workers to refer to the function of the Library of Congress (or Facebook's company processes) when they answer this question. They refer instead to their feelings about their own Facebook data.

We note other small clusters that have both answers and unusual interpretations in common, suggesting this phenomenon is highly local, rather than reflecting a single nexus of low fluency participation. Sometimes the clusters consist of survey pairs and triples. Cluster size may thus be related to how long the batch of 250 surveys takes to complete. A larger batch of HITs may reveal more extensive clusters of workers.

Table 5 shows a second cluster, again presented in submission order and pivoting on the "*Health Care*" response to AC2.

These six submissions show a markedly different interpretation of Q2 as a "how to" question rather than a policy question about social media archiving (as the question was interpreted by the responses in Figure 4). Respondents uniformly misinterpreted the question as, "How would you archive your Facebook content?" In past years, respondents had never interpreted Q2 as a 'how-to' question. Each of the six answers were copied from user help on Facebook's website. The first survey-taker, who submitted the survey before the others started, diverges in minor ways. The remaining five HITs were on workers' screens at overlapping times. The answers divide the Facebook help documentation among submissions. It is not clear whether this strategy represents one worker controlling multiple accounts, or several workers collaborating, each controlling multiple accounts.

Time to complete. Although work time may be confounded by tools and worker practices [23], it can still be a meaningful indication of how long it takes a good faith worker to complete the work. In 2013, the median time for workers to successfully complete this survey was about 10 minutes (the median value was 679 seconds). Some workers took longer, likely because their attention was divided or they were holding on to the survey to make sure their work wasn't rejected because they were working too fast. In 2013, only 14 out of 250 workers clocked in with work times over 30 minutes.

In 2022, the 28 workers who submitted acceptable surveys worked at about the same pace as workers had a decade earlier. The median work time was about 13 minutes. Six of the 28 workers had

Table 5. Crowdwork cluster, second example: possible collaboration and low fluency workers controlling more than one account

Q1: What types of things do you use the Internet for?	*AC1: Susie is looking for a job as a _____?*	*AC2: What's the name of the company where Greg and Homer work?*	*Q2: Would it be okay for a public institution to archive the contents of Facebook (including your stuff)? Why or why not?*	*Q3:Okay for FB to sell user data to Amason (Likert judgment)*	*Time spent on HIT (in seconds)*	*Accept time*
FACEBOOK	EXCELENT	HEALTH CARE	You can archive or delete some of the content you share on Facebook, like posts, photos and videos, directly from your News Feed or your timeline. From each post, you can choose to: Move to Archive. When you move content to your Archive, it'll only be visible to you.	6 - agree	640 (~11 min)	13:06:50
Downloading files.	NICE	HELTH CARE	When you archive a Facebook Group, y	5 - slightly agree	1101 (~19 min)	13:49:52
Downloading files.	NICE	HELTH CARE	you won't be able to create posts or like or add	5 - slightly agree	1294 (~22 min)	13:54:30
Downloading files.	GOOD	HALTH CARE	Can you archive Facebook content? You can archive or delete some of the content you share on Facebook, like posts, photos and vid"	4 - neutral	2086 (~35 min)	13:55:01
Downloading files.	NICE	HALTH CARE	You can archive or delete some of the content you share on Facebook, like posts, photos and videos, directly from your News Feed or your timeline. From each post, you can choose to: Move to Archive. When you move content to your Archive, it'll only be visible to you.	1 - strongly disagree	1748 (~29 min)	14:06:20
Downloading files.	NICE	HALTH CARE	Content that is defamatory, damaging, private, infringing, or otherwise unlawful can be removed from social media. There are a variety of tactics, both legal and non-legal, that are effective in securing the removal of harmful content.Jun 25, 2021	1 - strongly disagree	1659 (~28 min)	14:23:27

completion times of over 30 minutes. Of these six, three showed signs of reduced fluency, mostly in their open-ended answers. This increase in work time may not be strategic, but rather it may reflect how long the survey takes if a worker is not fully fluent in English.

We included the completion times and screen times for Tables 4 and 5. In both work clusters, the first submission has a work time close to the median, while others are either very long or much shorter than the median; both are signs that something is afoot [39]. Might a lead worker who is more proficient in English and more experienced on the platform be providing model answers for others doing the HIT?

5 Discussion

Given the power and utility of crowd survey platforms, we have no desire to abandon them. We expect the same type of non-malicious low fluency work to occur on other platforms too. Workers are often aware of other online venues to take surveys for pay and may have a multi-platform strategy to increase their earnings for this type of work [26]. New workers who make or watch introductory videos on YouTube often start with survey HITs. More advanced videos discuss tools and techniques (e.g., identifying attention checks and ways to identify and catch more desirable HITs). These videos are a good basis for understanding a worker's perspective, especially if you imagine low fluency workers receiving this advice in a language in which they are proficient. Besides explaining basic strategies, videos also direct workers to established worker forums, another venue for understanding an essentially invisible workforce whose practices aren't always obvious to requesters [36] [41].

Because we want to continue using these platforms, it is important to take a non-adversarial stance to good faith workers, even those we scrutinize in this paper. We also assume that providers like Amazon are only willing and able to do a certain amount to limit participation to the most qualified workers, especially if most requesters are satisfied with the work as it stands. Although there are intermediaries available to vet workers (e.g. Cloudresearch or CrowdFlower), it isn't clear that these platforms are addressing the problem of ambiguously capable workers. After all, some tasks (e.g. drawing a bounding box around an image) only require that workers understand the directions, while others, like content moderation, may demand situated cultural sensitivity. Still others, such as translation HITs, may take advantage of multilingualism and diverse experience. Crowdsourced tasks, like workers and requesters, exist along a complicated continuum. Furthermore, prequalified workers may not devote the same level of attention to the actual work they are qualified to do [30][37].

Are pandemic-era lockdowns responsible for the sharp uptick in unacceptable HIT performance on surveys? After all, with more

people staying at home (either by government policy or personal concern for safety), AMT provides a temporary income stream, albeit often not a sustainable salary in the US [36]. While lockdowns might be part of the problem, earlier signs of language proficiency issues have been documented [14].

In 2019, Amazon acknowledged worker identity fraud on Mechanical Turk [3]. By 2020, Amazon took action to quell this form of bad faith work [4]. But workers or intermediaries have evidently found a way around these measures. Although workers must present standard forms of identification used to get a job in the US—a social security number or other proof of eligibility—during pandemic work-from-home periods, workers in other countries developed ingenious workarounds and infrastructures to meet eligibility requirements to work in the US (e.g., [17]).

Are the good-faith efforts of low English proficiency workers influencing some published research results? Without indicting specific publications, it is easy to see how this work may be skewing survey research in many different Balkanized disciplines. But researchers mostly aiming to collect quantitative data may not notice changes in the platform.

6 Conclusion and Future Work

Like other researchers, we were surprised by the dramatic growth of unusable survey results on Mechanical Turk—from around 2% in 2013 to a little shy of 90% in 2022, compared using a stable set of questions and vetting mechanism. In this paper, we showed the effects of this phenomenon on Likert scale data and explored the sources of this work using qualitative methods. We discovered most of the unusable responses weren't from people working with malicious intent, but rather people with insufficient language and cultural fluency to give us meaningful results. We examined the work to distinguish and characterize six types:

Bad faith work. *Worker submits random responses or an incomplete survey*. Past findings suggest that much of the potential bad faith is weeded out by appropriate participation requirements [32]. The work currently perceived as bad faith often appears more random than it is. Language proficiency issues are at the root of some high-effort, low utility responses. Although actual bad faith work has presented problems that require tuning participation requirements, it now seems far less common than the other types of unusable work. Apparent bad faith work also may be exacerbated by requester-side factors like HIT design for different work environments [13] or clarity [15].

Ultra-low fluency work. *Worker submits responses copied from a SERP or web page using HIT text as a query*. Because copying text can be a time-consuming practice for workers if it is done without software assistance, we suppose copying is done by workers with little understanding of the HIT's language. Without open-ended questions, this work may be difficult to detect: some workers who responded to open-ended questions with SERP snippets answered both attention checks correctly.

Low fluency work. *Worker submits responses that have probably required assistance from translation software*. This work may meet the requester's requirements, depending on why the requester wanted US-based workers. Again, without open-ended questions, this work is not only difficult to detect, but also may require a judgment call. Like ultra-low fluency work, workers may answer the attention questions correctly. The survey responses yield more information than satisficing work, but it's never certain that the worker understood the nuances of the questions.

Low fluency collaborative or duplicative work. *Workers submit overlapping distinctive responses to pivotal questions*. Similar to the findings of Checco et al. [7], some responses appear to be the efforts of people working together to understand the survey. We also see signs of individuals controlling multiple worker accounts. Lead workers with greater English proficiency may also model answers for colleagues.

Satisficing work. *Worker submits minimal responses and occasionally skips questions*. This type of work is perhaps the most frustrating to qualitative researchers who appreciated the type of diverse participation Mechanical Turk offered [5][31][33]. This is work that might be improved by offering incentives for better quality participation [29], although further incentives might be what reveals shortfalls in language proficiency. Survey platform burnout can turn workers into satisficers [18].

Good faith, high engagement work. *Worker submits complete and engaged responses*. This is the type of work that was common in the early days of the platform; it's what drew many of us to run studies using AMT. High engagement work can be the result of something the requester did right (incremental incentives, known-good workers, interesting topic, a well-designed survey), but it may also be that the worker isn't burned out on surveys yet. New workers who take surveys can be highly motivated to venture opinions or information about themselves.

Given all these categories, there are other confounding factors like worker farms (which may be behind the collaborative efforts), VPNs that enable workers to conceal their location [9], and cloud computing platforms such as Oracle Public Cloud. Some survey platforms seem to blacklist common VPNs and cloud computing services to mitigate fraud. Less malicious environmental factors may come into play too [13]. E.g., some workers may be taking surveys on smart phones or tablets regardless of the instructions.

Efforts to develop workarounds include filtering methods to prequalify respondents as Dupuis et al. are testing [12]; pressure campaigns that demand platforms like AMT "do better" at ensuring workers are who they say they are [3]; and cleaning methods that accept low fluency work and recruit a larger number of respondents to compensate as principled ways to clean the data improve. In future work, we plan to develop algorithmic methods of classifying responses according to meta-attributes like language proficiency and signals that the work is being done in good faith.

For some tasks, language proficiency or cultural context will be more important, Other tasks may be able to take advantage of this inadvertent diversity (or at least not be hampered by it) [22].

As part of our future work, we plan to conduct experiments to replicate existing non-survey work such as labeling and relevance judgement to determine the effect of low language fluency work on other reliability metrics such as inter-rater agreement and comparison with gold set values. We also plan to verify whether other for-pay online survey research platforms like Prolific and

Dynata are experiencing a similar increase in low fluency work. We might expect a larger distinction to be between online surveys and more traditional means of survey data collection [19] and for survey-specific online sites to be even more apt to attract satisficers than AMT [42].

Although we would like to rely on the online crowdwork platform's providers like Amazon to regulate worker identity—and for US workers to earn an appropriate wage for their efforts—we fear that this may not be a reasonable expectation. Crowdwork platforms may feel they already do enough, especially if low-fluency work elicits few complaints from requesters. Further investigations of good-faith, low fluency work may help us detect and evaluate work in a way that is appropriate and fair.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1816923. We would like to thank the anonymous workers who contributed this study.

REFERENCES

[1] Douglas J. Ahler, Carolyn E. Rous, Gaurav Sood. 2020. The Micro-Task Market for Lemons: Data Quality on Amazon's Mechanical Turk.

[2] Omar Alonso, Catherine C. Marshall, and Marc Najork. 2015. Debugging a Crowdsourced Task with Low Inter-Rater Agreement. In *Proc. JCDL 2015*.

[3] Amazon Mechanical Turk. 2019. MTurk Worker Quality and Identity. Blog post dated 25 March 2019. https://blog.mturk.com/mturk-worker-identity-and-task-quality-d3be46d83d0d

[4] Amazon Mechanical Turk. 2020. Important updates on MTurk marketplace integrity, Worker identity and Requester tools to manage task quality. Blog post dated 20 March 2020.

[5] F. Bentley, N. Daskalova, and B. White. 2017 Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys. *Proc. CHI EA '17*. ACM, New York, NY, USA, 1092-1099.

[6] Erin M. Buchanan and John E. Scofield. 2018. Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods 50*, 2586–2596.

[7] Checco, A., Bates, J., & Demartini, G. (2018). All That Glitters Is Gold — An Attack Scheme on Gold Questions in Crowdsourcing. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 6(1), 2-11. https://doi.org/10.1609/hcomp.v6i1.13332

[8] Michael Chmielewski and Sarah C. Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science 11*, 4 (2020), 464–473.

[9] S.A. Dennis, B.M. Goodson, and C. Pearson. 2018. MTurk workers' use of low-cost "virtual private servers" to circumvent screening methods. Research note. https://doi.org/10.2139/ssrn.3233954.

[10] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proc. WSDM 2018*. https://doi.org/10.1145/3159652.3159661.

[11] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system? screening mechanical turk workers. In *Proc. CHI '10*. ACM, New York, NY, USA, 2399–2402.

[12] Marc J. Dupuis, Karen Renaud, and Rosalind Searle. 2022. Crowdsourcing Quality Concerns: An Examination of Amazon's Mechanical Turk. In *Proc. SIGITE '22*, ACM, New York. https://doi.org/10.1145/3537674.3555783.

[13] Ujwal Gadiraju, Alessandro Checco, Gupta, N., & Demartini, G. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. Proc. Interactive, Mobile, Wearable and Ubiquitous Tech., *1* (3), 1-29.

[14] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In *Proc. CHI'15*. ACM, 1631-1640.

[15] Ujwal Gadiraju, Jie Yang and Alessandro Bozzon. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. *Proc. HT'17*, 5-14.

[16] Mary L. Gray. Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The Crowd is a Collaborative Network. *CSCW '16*, February 27-March 02, 2016, San Francisco, CA, USA. ACM, New York.

[17] Guidance on the Democratic People's Republic of Korea Information Technology Workers. 2022. US Gov't 16 May 2022 https://home.treasury.gov/system/files/126/20220516_dprk_it_worker_advisory.pdf.

[18] Tyler Hamby and Wyn Taylor. 2016. Survey Satisficing Inflates Reliability and Validity Measures: An Experimental Comparison of College and Amazon Mechanical Turk Samples. *Educ Psychol Meas. 6*(6), 912–932.

[19] Eszter Hargittai and Aaron Shaw. 2020. Comparing Internet Experiences and Prosociality in Amazon Mechanical Turk and Population-Based Survey Samples. Socius, 6 (1), 1–11.

[20] D.J. Hauser and N. Schwarz. 2015. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods 48*, 400–407.

[21] Jakobsson, M. 2009. Experimenting on Mechanical Turk: 5 How Tos. *ITWorld*, September 3, 2009.

[22] Shivani Kapania, Ding Wang, and Alex Taylor. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In Proc. CHI '23. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3580645.

[23] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P. Bigham. 2018, Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers. In *The Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018)*. AAAI Press, pp 70-78.

[24] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. 2020. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods 8*, 614–629. doi:10.1017/psrm.2020.6.

[25] Kittur, A., Chi, E., Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proc. CHI'08*, 453-456.

[26] Catherine C. Marshall and Frank M. Shipman. 2013. Experiences Surveying the Crowd: Reflections on Methods, Participation, and Reliability, *Proc. WebSci'13*, 234-243.

[27] Catherine C. Marshall and Frank M. Shipman. 2015. Exploring the Ownership and Persistent Value of Facebook Content. *Proc. CSCW'15*, ACM Press, NY.

[28] Catherine C. Marshall and Frank M. Shipman. 2017. Who Owns the Social Web? CACM 60, 5, 52-61.

[29] Winter Mason and Duncan J. Watts. 2009. Financial incentives and the "performance of crowds". *HCOMP '09*. ACM, New York, NY, USA, 77–85.

[30] Adam W. Meade and S. Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455.

[31] Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavioral Research 54*, 1643–1662.

[32] Eyal Peer, J. Vosgerau, and A. Acquisti. 2014. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods 46*, no. 4, pp. 1023–1031.

[33] Elissa M. Redmiles, Sean Kros, and Michelle L. Mazurek. 2019. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples.

[34] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI EA '10*. ACM, NY, USA, 2863–2872.

[35] Timothy J. Ryan. 2020. Fraudulent responses on Amazon Mechanical Turk: A Fresh Cautionary Tale. Blog post published 22 December 2020 on https://timryan.web.unc.edu.

[36] Shruti Sannon and Dan Cosley. 2019. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. In *CHI 2019*, ACM, New York.

[37] Antonios Saravanos, Stavros Zervoudakis, Dongnanzi Zheng, Neil Stott, Bohdan Hawryluk, Donatella Delfino. 2021. The Hidden Cost of Using Amazon Mechanical Turk for Research. *Lecture Notes in Computer Science 13094*. Springer. https://doi.org/10.1007/978-3-030-90238-4_12

[38] Frank M. Shipman and Catherine C. Marshall. 2020. Ownership, Privacy, and Control in the Wake of Cambridge Analytica. *Proc CHI 2020*. ACM, New York.

[39] Andie Storozuk, Marilyn Ashley, Veronic Delage, and Erin A. Maloney. 2020. Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks. *Quantitative Methods for Psychology* 16 (5), 472-481.

[40] Martine Van Selm and Nicholas W. Jankowski. 2006. Conducting Online Surveys. *Quality & Quantity 40*, 435–456.

[41] Alex C. Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork. *Proceedings of CSCW'19*, 1–28.

[42] Bingbing Zhang and Sherice Gearhart. 2020. Collecting Online Survey Data: A Comparison of Data Quality among a Commercial Panel and MTurk. Survey Practice 13 (1). https://doi.org/10.29115/SP-2020-0015.