Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Generative AI in Crowdwork for Web and Social Media Research: A Survey of Workers at Three Platforms

Evgenia Christoforou[1], **Gianluca Demartini**[2], **Jahna Otterbacher** [1,3]

[1]CYENS - Centre of Excellence, Nicosia, Cyprus
[2]The University of Queensland, Australia
[3] Open University of Cyprus, Nicosia, Cyprus

Abstract

Crowdsourcing plays an important role in Web and social media research - from data annotation, to online experiments and user surveys. With the emergence of Generative AI (GenAI), researchers are considering how models and tools such as GPT might replace crowdwork. Many have already evaluated GPT on annotation tasks. However, it is less clear how GenAI might impact other types of tasks, or to what extent crowdworkers have already incorporated it into their work processes. Thus, we asked crowdworkers directly regarding their use of GenAI, via a survey at two points in time, across three commercial platforms. We found evidence that workers' self-reported use of GenAI did not change over time, but rather, was strongly correlated to the platform in which they operate, with MTurk workers using GenAI much more often than those operating at Clickworker and Prolific. As most respondents reported that survey completion is their "usual type of task," we discuss the implications of the use of GenAI in user surveys, via specific examples of ICWSM research.

Introduction

For more than ten years, paid micro-task crowdsourcing has served as a means to collect data annotations and to perform research studies with human subjects at scale. Traditionally, researchers have relied on crowdsourcing for quality data, when it comes to tasks on which humans outperform machines. The recent advent of GenAI has the potential to disrupt current crowdsourcing practices, since powerful large language models (LLMs), such as those in the Generative Pre-trained Transformer (GPT) family, have become readily available to researchers, developers, and laypersons alike.

It is obvious that GenAI will change – if not totally replace – crowdwork for *data annotation tasks*. In particular, OpenAI actively encourages the use of its GPT API for text analysis; its Prompt Engineering guidelines[1] demonstrate the creation of text classifiers, providing a specific example of sentiment detection on tweets. As expected, there has been a surge of research investigating how GPT-generated data annotations compare to those generated by humans. For example, using a TREC dataset, Thomas et al. (2023) found that relevance labels generated by GPT-4 are comparable to those of humans, suggesting that the LLM can understand searcher preferences. In a similar vein, Gilardi, Alizadeh, and Kubli (2023) demonstrated that GPT-3.5 was more accurate and consistent than trained political science students in labeling politically-oriented tweets on a number of dimensions (relevance, stance and frame detection). Others have reported promising results on hate detection (Huang, Kwak, and An 2023), sentiment detection and bot detection (Zhu et al. 2023), to name but a few studies and tasks.

It is clear that GenAI may increasingly be used by researchers as a substitute for crowdwork, particularly in data annotation. But in cases where researchers continue to crowdsource,[2] we must also ask if and how crowdworkers incorporate GenAI into their work processes. It is not obvious how to control if and how much workers on commercial platforms make use of GenAI as a co-pilot during their task completion efforts. To address this, Veselovsky, Ribeiro, and West (2023) looked at the use of LLMs by crowdworkers completing a text summarization task on Amazon MTurk. Using keystroke analysis, as well as synthetic text detection, they estimated that around 33-46% of crowdworkers used LLMs when completing the task. We add to the emerging discussion surrounding if and how crowdworkers use GenAI, using an alternative approach – we ask workers directly, via a survey at three commercial platforms. Currently, we address the following research questions (RQs):

- Do crowdworkers on commercial platforms use GenAI tools in their tasks, even when not asked to do so?
- Is there an increasing tendency to use GenAI over time?
- What are the implications for ICWSM research?

Crowdwork in ICWSM Research

Before presenting our methodology, it is useful to reflect briefly on how the ICWSM community has used crowdsourcing over the years. Via the AAAI ICWSM online proceedings, we retrieved papers containing at least one of the following keywords: *crowd(work), crowdsourc(ing), worker, Mechanical Turk, Crowdflower, Appen, Figure Eight, Clickworker, Prolific, Microworkers, Rapidworkers*. We identified 27 papers in which the authors conducted, first-hand,

[1]https://platform.openai.com/docs/guides/prompt-engineering

[2]The question of when, where and why human-only work is required is an interesting one, which we must leave for future work.

a crowdsourcing task on a commercial platform.[3] Of the 27 studies, the platforms used were: Mechanical Turk (18), Crowdflower/Appen/FigureEight (7), other (2). While 13 studies described data annotation tasks, the others included user surveys, online experiments, and content creation tasks, usually in combination with one another. Next, we provide examples of how ICWSM researchers used commercial crowdsourcing to solicit *human* perceptions and experiences with/within Web and social media. While not an exhaustive presentation, it will allow us to explore the implications of our findings concerning crowdworkers' self-reported usage of GenAI on tasks beyond data annotation, which as mentioned, represents the lion's share of research to date.

User surveys. (1) *Assessing motivations.* Heckner, Heilemann, and Wolff (2009) used MTurk to execute a user survey, to learn why people participate in social tagging systems. They noted that crowdsourcing is "comparatively cheap," allowing them to create a survey "incorporating more than one hundred test subjects," who were "real users" of four systems (Flickr, Youtube, Delicious and Connotea). (2) *Understanding user behavior and expression.* De Choudhury et al. (2013) used an MTurk survey to administer a standardized clinical screening for depression. They then invited participants to share their Twitter profile, as well as access to their Twitter data, for research purposes. In a similar spirit, Boyd et al. (2015) aimed to use language analysis to understand human values. Via an MTurk survey, they asked participants to complete two essays describing their values and behaviors, as well as to complete a standardized test to assess their values. (3) *Pre-screening for future tasks.* Surveys have also been used as a means to collect data about user demographics, behaviors and views, as a means to create a pool of participants who researchers may then invite to participate in **online experiments**, which are also often conducted via MTurk (Rogstadius et al. 2011; Celis, Krafft, and Kobe 2016; Ye, You, and Robert Jr 2017; Tausczik and Boons 2018; Burghardt, Hogg, and Lerman 2018). (4) *Post-task reflections.* Likewise, user surveys can be administered following the execution of another crowd task (e.g., post-annotation (Biel and Gatica-Perez 2012)). (5) *Improving platform features.* Another example is the work by Rokicki, Trattner, and Herder (2018), who conducted a survey at Crowdflower, in which participants were shown a set of online recipes and were asked to estimate their healthiness and caloric content. Here, the goal was to provide insights as to features that improve user understanding on the recipes.

Content creation or evaluation. (1) *Generating baseline content.* Mukherjee et al. (2013) studied the phenomenon of deceptive opinion spamming, and its detection, crowdsourcing a set of fake hotel reviews from MTurk workers. (2) *Quality control.* Chua and Asur (2013) asked MTurk workers to evaluate automatically-generated summaries of events, describing a need to perform "qualitative assessment from human readers."

[3]We did not consider papers describing secondary data analyses of crowdsourced data or crowdsourcing via other means, despite that GenAI will impact such cases as well, likely in similar ways.

	Prolific	MTurk	Clickworker
USA	**199** (200) **200** (200)	**194** (200) **194** (200)	**183** (200) **141** (200)
UK	**99** (100) **100** (100)	- -	**83** (100) **82** (100)
EU (GER, ITA, FRA, ESP)	**200** (200) **200** (200)	- -	**189** (200) **191** (200)
India	- -	**196** (200) **187** (198)	- -

Table 1: Total # responses (in parenthesis) and # considered after cleaning, T1 (top) and T2 (bottom).

Methodology

Our questionnaire captured the following: (1) experience and (2) motivation of participants, (3) platforms used and types of tasks performed, (4) experience with GenAI such as ChatGPT, (5) use / intended use of GenAI, (6) completion of tasks for identifying machine generated content, (7) opinion on the impact of GenAI on crowdwork more generally. No personal or identifying information was collected. To answer our current RQs, we focus only on analyzing the data captured concerning workers' use of GenAI and how they believe it might impact the usual tasks they perform.

We posted our questionnaire to three platforms: (1) Prolific, (2) MTurk and (3) Clickworker, which are open to academic researchers. Since this is no longer the case for Appen (formerly Crowdflower/Figure Eight), we included two newer platforms, along with MTurk, which are becoming popular with researchers. We aimed for a balanced number of responses gender-wise. Additionally, according to the platforms[4,5,6] and documentation on crowd populations (Posch et al. 2022; Difallah, Filatova, and Ipeirotis 2018), we aimed at workers residing in countries with a known presence at each platform. Apart from specifying the country of residence and guaranteeing a balanced gender sample, we did not impose further restrictions on the workers who could participate. We collected responses at two points in time (T1, T2): May and December 2023.

Table 1 details the responses collected in parentheses, and in bold, the final number after cleaning. We used an external link to administer our task, to provide the same survey across platforms; the cleaning process removed responses where the worker failed to prove their identity in the platform (i.e., a valid crowdworker id). Workers were first informed of the study objectives, and their right to withdraw at any point. The study received ethical approval from the Cyprus National Bioethics Committee and workers were rewarded fairly according to the respective platform's instructions, respecting the average hourly salary per country.[7]

[4]https://researcher-help.prolific.co/hc/en-gb/articles/360009220833-Who-are-the-participants-on-Prolific-#heading-0

[5]https://www.clickworker.com/clickworker-crowd/

[6]https://www.mturk.com/help

[7]Dataset is available at: https://zenodo.org/records/10886052

	Prolific	MTurk	Clickworker
Surveys	89% 87%	52% 46%	67% 59%
Visual object identification	2% 2%	17% 18%	7% 7%
Text annotation	2% 3%	17% 23%	7% 11%
Find information about an entity	1% 3%	9% 7%	7% 7%
Create content	3% 3%	2% 4%	5% 6%
Access a link to read content	1% 1%	1% 1%	4% 4%
Other task	2% 1%	2% 1%	3% 6%

Table 2: Workers' "usual task" (T1/T2 [top/bottom]).

	ALL	USA	India	UK	EU
Prolific	47.8% 62.2%	44.7% 57.0%	- -	34.3% 47.0%	57.5% 75.0%
MTurk	88.7% 83.6%	94.3% 90.4%	83.1% 76.4%	- -	- -
Clickworker	46.4% 58.7%	48.6% 64.5%	- -	48.1% 48.8%	43.4% 58.6%

Table 3: Workers reporting use of AI chatbots in everyday life, by platform, country and T1/T2 [top/bottom].

	ALL	USA	India	UK	EU
Prolific	13.1% 13.4%	19.0% 14.0%	- -	9.0 % 10.0%	9.0% 14.5%
MTurk	80.3% 73.2%	94.3% 86.2%	66.3% 59.4%	- -	- -
Clickworker	20.7% 15.0%	27.9% 20.6%	- -	16.9% 11.0%	15.3% 12.6%

Table 4: Workers reporting self-initiated use of AI chatbots in tasks, by platform, country and T1/T2 [top/bottom].

Findings

Crowdworkers' Experience and Usual Tasks

We asked participants to describe their experience on the respective crowdsourcing platform, reporting their total number of completed tasks. This was to ensure that our pool of respondents adequately captured workers with a reasonable level of experience. We performed a check on two categories at the extremes: "newcomers" are those having completed few tasks ("between 0-10 tasks"), and with a short timewise experience ("first time doing crowdsourcing"), while "legacy" crowdworkers are those with more than a year of experience and more than 100 tasks completed. Our sample contained relatively few newcomers at each platform and at T1/T2 (*Prolific*: 16.7% / 14.6%, *MTurk*: 3.8% / 10.9%, *Clickworker*: 10.8% / 15.9%). We were also able to capture a good number of legacy workers: (*Prolific*: 49.0% / 45.4%, *MTurk*: 17.2% / 22.1%, *Clickworker*: 35.2% / 51.6%).

We also asked participants to report their "usual type of task" at the respective platform. This was a closed-formed question, in which they could choose one of six example tasks or "other." As can be seen in Table 2, *surveys* represent the typical task for most of our respondents. This is particularly true of those who engage with Prolific, where nearly 90% – at both points in time – chose "surveys" as their typical task. It is clear that workers at MTurk (and to a lesser extent Clickworker), engaged with a variety of other tasks as well (e.g., visual object identification, text annotation).

Crowdworkers' Use of GenAI

To establish that crowdworkers are aware of this disruptive new technology, we first posed the question: *Have you ever used an AI Chatbot, like ChatGPT, to find information in your everyday life?* Table 3 presents the percentage of workers in each platform and country that use ChatGPT in everyday life, at T1 and T2. MTurk workers, at both points in time, are more likely than others to report using AI chatbots. Considering workers across all locations, a chi-square test confirms a significant platform effect at both T1 ($X^2(3, N = 1298) = 198.5, p < 0.001$) and T2 ($X^2(3, N = 1298) = 66.3, p < 0.001$). More Prolific ($z = 4.57, p < 0.001$) and Clickworkers ($z = 3.62, p < 0.001$) participants reported using AI chatbots at T2 versus T1. However, there was no significant difference in the proportion of MTurk workers reporting using the technology, from T1 to T2.

Table 4 presents the distribution of responses to the question: *Have you ever used an AI chatbot, like ChatGPT, to help you complete a crowdsourcing task without being instructed by the task to do it?)*. Here, there is also strong evidence of a platform effect, with MTurk workers being much more likely to report using AI, as compared to others, at both T1 ($X^2(3, N = 1298) = 496.7, p < 0.001$) and T2 ($X^2(3, N = 1298) = 436.8, p < 0.001$). One might expect to observe more self-reported use of AI in T2, but this also does not appear to be the case. The proportion of MTurk workers reporting use at T2 is actually less than at T1 ($z = 2.33, p < 0.01$), but the differences are not statistically significant for responses from the other two platforms.

GenAI and User Surveys

Clearly, GenAI may end up replacing (human) crowdworkers in a number of tasks, especially those related to text annotation and image analysis (Thomas et al. 2023). However, it is less clear how other tasks, such as user surveys, may be impacted. To shed light on what workers may be thinking as to the specific impact of GenAI on surveys, we considered the free-text responses to the follow-up question: *Do you think that Artificial Intelligence (AI) chatbots or image generators will have a positive impact on the practice of crowdsourcing? Please explain your thoughts on this.* In particular, we examined the 76 responses containing the word "survey(s)" (65 from Prolific, 9 from MTurk, 2 from Clickworker). Workers highlighted two reasons why GenAI might result in an *increase* in available tasks – 1) increased interest and research on GenAI, 2) GenAI will help researchers quickly generate more surveys of a high quality. These views are represented by the following example responses:

–In the case of Prolific, AI chatbots have seemed to increase the number of surveys because people are incredibly interested in researching them. (Prolific, T1)
–People who conduct surveys will be faster at generating content and will be more original. (Prolific, T2).

In contrast, several reasons for a possible *decrease* in tasks were shared by workers. These included – 1) increased cheating and abuse by workers, 2) decrease in data quality / generic data, 3) answers that are detached from reality, 4) answers that lack human experience and/or emotion. Such reasons were expressed in the following example responses:

–People want real answers from humans and not just simulated answers from a bot. (MTurk, T1)
–AI is not feel human emotion so some survey which want human emotion to solve problem so that case AI not work. (MTurk, T1)
–i think it could be negative since they are conducting surveys on a "crowd" of people. i don't think the AI has the knowledge about the real world. (Prolific, T1)
–I think that AI will generally reduce tasks overall that require human intervention besides academic surveys. I also think that people could abuse such chatbots (MTurk, T2)
–Negative impact because soon AI will be used to simulate answers for different persons hence reducing the amount of available surveys. (Prolific, T1)

Conclusions

To date, most work considering the impact of GenAI on crowdsourcing has focused on evaluating its use on *data annotation* tasks. Beyond annotation, the ICWSM community has used commercial crowdsourcing in a variety of ways, including for the administering of user surveys and online experiments, in an effort to understand "real users'" behaviors within and experiences with networked media. Thus, to understand the impact of GenAI on our research, it is necessary to ask whether crowdworkers themselves are incorporating GenAI into their tasks, in ways that could stand to compromise the authenticity of the collected data.

We asked crowdworkers directly, via a survey, if they have used AI chatbots in their tasks, on their own volition. Thus, in contrast to Veselovsky, Ribeiro, and West (2023), we cannot report on what they actually do during a task, but rather, what they say that they do. Through their self-reports, we found evidence of a strong platform effect, in terms of workers' use of AI chatbots in general, and their use of it on task. In particular, Amazon Mechanical Turk respondents were more likely to use AI chatbots in everyday life, and were also more likely to report having used it while crowdworking, without having been told to do so. This was true at both points in time, September and December 2023, a period of time during which the public was increasingly adopting GenAI applications such at ChatGPT.[8]

Our respondents overwhelmingly indicated that *surveys* represent their "usual" type of tasks that they complete. This raises a red flag for researchers of the Web and social media. In our review of ICWSM work over the years, we found that many relied on MTurk, the platform with the most self-reported use of GenAI, to execute surveys and experiments. In fact, in the analysis of respondents' free-text responses, in which they described their beliefs as to the impact of GenAI on the number of available tasks for them to complete, it is clear that they are well aware of the various threats that GenAI poses to data quality and integrity.

Based on our findings, researchers must think carefully about their choice of platform for crowdsourcing, especially for studies that aim to assess genuine human insight and real-world experience. It is clear that they will need to explicitly state whether or not using GenAI is acceptable on task. At the same time, researchers should also keep informed on the platforms' policies towards the use of GenAI, which likely influence the work ethic – and the cross-platform differences we have observed. In short, it may be the case that commercial crowdsourcing platforms can no longer be considered by researchers as a go-to place to find a convenient and inexpensive pool of "real users" to recruit as study participants. Certainly, in light of GenAI, researchers will need to be (even more) diligent in considering the quality and validity of their collected data.

As with all research on social phenomena, our study has limitations and potential negative effects, which must be considered when interpreting and/or applying our findings. First, our study provides a snapshot of self-reports from workers at three platforms, at two points in time, and our study is cross-sectional (i.e., we have not followed the same workers across time). Secondly, while overwhelmingly, participants indicated that surveys are their "usual task," we cannot make specific claims as to how they use GenAI on this particular class of tasks. Nonetheless, we believe that there is cause for concern, in terms of GenAI impacting tasks beyond annotation, given that participants report using it on their own initiative (i.e., when not instructed to do so). Finally, there is a potential negative societal impact that could result from the publication of our findings – researchers may end up reducing their use of crowdsourcing, resulting in a loss of opportunities for workers operating at the three platforms studied. However, as our participants themselves expressed in their free-text responses, researchers will ultimately be influenced by the quality of the results they obtain from crowd work.

In conclusion, we provided a first look at the self-reported use of GenAI at three micro-task crowdworking platforms. Given the rapidly evolving area of GenAI, and its potential to disrupt work processes, there will be a constant need to evaluate the use of GenAI via multiple methodologies (e.g., user experiments, system capability evaluations, worker self-reports) across time (Christoforou, Barlas, and Otterbacher 2021). Such studies can aid researchers in understanding the impact of GenAI, as to preserve the quality and validity of their scientific data.

[8]https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Acknowledgments

This project has received funding from the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0421/0360 (KeepA(n)I), the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 739578 (RISE), the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy, and by the Australian Research Council (ARC) Training Centre for Information Resilience (Grant No. IC200100022).

References

Biel, J.-I.; and Gatica-Perez, D. 2012. The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, 407–410.

Boyd, R.; Wilson, S.; Pennebaker, J.; Kosinski, M.; Stillwell, D.; and Mihalcea, R. 2015. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 31–40.

Burghardt, K.; Hogg, T.; and Lerman, K. 2018. Quantifying the impact of cognitive biases in question-answering systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Celis, L. E.; Krafft, P.; and Kobe, N. 2016. Sequential voting promotes collective discovery in social recommendation systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, 42–51.

Christoforou, E.; Barlas, P.; and Otterbacher, J. 2021. It's about time: A view of crowdsourced data before and during the pandemic. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14.

Chua, F.; and Asur, S. 2013. Automatic summarization of events from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 81–90.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 128–137.

Difallah, D.; Filatova, E.; and Ipeirotis, P. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 135–143.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

Heckner, M.; Heilemann, M.; and Wolff, C. 2009. Personal information management vs. resource sharing: Towards a model of information behavior in social tagging systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, 42–49.

Huang, F.; Kwak, H.; and An, J. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, 294–297.

Mukherjee, A.; Venkataraman, V.; Liu, B.; and Glance, N. 2013. What yelp fake review filter might be doing? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 409–418.

Posch, L.; Bleier, A.; Flöck, F.; Lechner, C. M.; Kinder-Kurlanda, K.; Helic, D.; and Strohmaier, M. 2022. Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics. *Human Computation*, 9(1): 22–57.

Rogstadius, J.; Kostakos, V.; Kittur, A.; Smus, B.; Laredo, J.; and Vukovic, M. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 321–328.

Rokicki, M.; Trattner, C.; and Herder, E. 2018. The impact of recipe features, social cues and demographics on estimating the healthiness of online recipes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Tausczik, Y.; and Boons, M. 2018. Distributed knowledge in crowds: Crowd performance on hidden profile tasks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Thomas, P.; Spielman, S.; Craswell, N.; and Mitra, B. 2023. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*.

Veselovsky, V.; Ribeiro, M. H.; and West, R. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv preprint arXiv:2306.07899*.

Ye, T.; You, S.; and Robert Jr, L. 2017. When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 327–336.

Zhu, Y.; Zhang, P.; Haq, E.-U.; Hui, P.; and Tyson, G. 2023. Can chatGPT reproduce human-generated labels? A study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

Paper Checklist

1. (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Answering the posed RQs does not require us to violate user privacy or treat people unfairly. In contrast, answering the RQs will help us understand how a key research method used in the community is likely to be impacted. It also helps to keep some transparency between participants (crowdworkers in this case) and researchers (data requesters).
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No. We feel that there is really only one potential negative impact of this work, which could be decreased interest in creating crowd tasks. However, as discussed in the final section, that will primarily be driven by the quality of data that researchers will be able to obtain; not so much by our exploratory study.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? This is a four page paper. Thus, we will have to release that with the dataset, which will be made available upon acceptance of the paper.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? N/A
   (b) Did you mention the license of the assets? N/A
   (c) Did you include any new assets in the supplemental material or as a URL? We have provided a sample of the data collected from the Clickworker platform.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Since this is a four page paper, we will include the full questionnaire and consent procedure along with the dataset release. In the body of the paper, we have indicated that participants were explained the objectives of the study and that they could leave at any time without having their data captured.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? We indicated that the questionnaire is anonymous. We have only the worker IDs for the respective platfrom, which will not be included in the data release.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? This is a four page paper. Thus, we will have to release that with the dataset, which will be made available upon acceptance of the paper.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? This is a four page paper. Thus, we will have to release that with the dataset, which will be made available upon acceptance of the paper.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? This is a four page paper. Thus, we will have to release that with the dataset, which will be made available upon acceptance of the paper. In addition, we have analyzed in this paper, only a subset of questions that were on our questionnaire. Those questions have been presented in full.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Our research protocol has received ethical approval from our institution. We have mentioned that and will provide the protocol number upon acceptance of the paper.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? This is a four page paper. Thus, we will have to release the details of how fair payments were calculated with the dataset, which will be made available upon acceptance of the paper. In the body of the paper, we have confirmed that "crowdworkers were rewarded fairly according to the respective platform's instructions, respecting the average hourly salary per country."
   (d) Did you discuss how data is stored, shared, and deidentified? This is a four page paper. Thus, we will have to release that with the dataset, which will be made available upon acceptance of the paper.