Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing

Jonathan K. Kummerfeld
Computer Science & Engineering
University of Michigan, Ann Arbor
jkummerf@umich.edu

Abstract

Extensive work has argued in favour of paying crowd workers a wage that is at least equivalent to the U.S. federal minimum wage. Meanwhile, research on collecting high quality annotations suggests using a qualification that requires workers to have previously completed a certain number of tasks. If most requesters who pay fairly require workers to have completed a large number of tasks already then workers need to complete a substantial amount of poorly paid work before they can earn a fair wage. Through analysis of worker discussions and guidance for researchers, we estimate that workers spend approximately 2.25 months of full time effort on poorly paid tasks in order to get the qualifications needed for better paid tasks. We discuss alternatives to this qualification and conduct a study of the correlation between qualifications and work quality on two NLP tasks. We find that it is possible to reduce the burden on workers while still collecting high quality data.

1 Introduction

Workers using Amazon Mechanical Turk earn a median wage of $2.54 an hour (Hara et al., 2018), far below the U.S.-federal minimum wage of $7.25. Many researchers pay workers a higher wage, estimating the time spent on a task and giving bonuses when the time required is higher than expected. At the same time, researchers try to maintain the quality of work completed using a variety of methods (Mitra et al., 2015). One common approach, used by 19% of tasks (HITs) on the platform (Hara et al., 2018), is to restrict tasks to workers who have had a certain number of HITs approved. Tasks with this restriction have a median wage of $4.14 an hour, far above the overall average. If most high paying requesters use this restriction it means workers need to do a substantial amount of low paid "Qualification Labour": work to achieve the qualifications necessary for fairly paid tasks. These tasks may also be particularly unpleasant work that more experienced workers are unwilling to do, e.g., they might involve unsavoury content.

This paper is the first to identify the qualification labour issue and explore it. We study norms around the setting of the qualification and the effort workers put in to achieve common milestones. 5,000 accepted tasks, a common requirement, takes over 2 months of effort. We consider several ways to address the issue, and study the work quality of groups with different qualifications.[1] Using two tasks, coreference resolution and sentiment analysis, we find that high quality annotations can be collected with a lower threshold, though there are task dependent patterns.

2 Background and Related Work

Crowd work involves large groups of workers doing small paid tasks, known as Human Intelligence Tasks (HITs). Services such as Amazon Mechanical Turk provide a marketplace to connect workers with requesters. Requesters create tasks, workers choose which tasks to do, then either complete them or return them. Requesters approve or reject the completed work. Tasks can be restricted to workers with certain qualifications, e.g. location. Amazon tracks some statistics that can be used as qualifications. This work focuses on (1) the total number of approved HITs a worker has, and (2) the percentage of their HITs that were accepted.

Since the earliest uses of crowd work in NLP, there has been work discussing issues such as poor wages and the lack of worker rights (Fort et al., 2011). These have also been discussed in the Human-Computer Interaction research community (Bederson and Quinn, 2011; Hara et al., 2018). There has been work on proposing guidelines for requesters (Sabou et al., 2014), incorporating workers into the IRB process (Libuše Hannah Vepřek,

[1] Code for our experiments is attached to this paper

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 343–349

2020), and developing tools to help workers address the power imbalance in the online workplace (Irani and Silberman, 2013, 2016). Concurrent with this work, another study showed that crowdsourcing is being used more each year in NLP research, and there is limited awareness of the ethical issues in this type of work (Shmueli et al., 2021).

Prior work has considered hidden labour in the day-to-day work of the crowd (Hara et al., 2018). By observing a large set of workers, they measured time involved in searching for tasks, returned tasks, and breaks. Some of these issues have received additional attention, such as the wasted effort on tasks that are returned rather than completed (Han et al., 2019). While informative, those studies do not account for the hidden labour identified in this paper, which spans a long period and relates to worker qualifications.

Part of this work uses online discussion between workers to understand their work. Prior work has used a similar approach to understand the overall experience of crowd workers (Martin et al., 2014).

3 Norms for the Approved HITs Value

The value used as the Approved HITs threshold is rarely reported in prior work. Three recent papers specify a 1,000 HIT threshold (Vandenhof, 2019; Oppenlaender et al., 2020; Whiting et al., 2019). Outside of Computer Science, advice in articles (Young and Young, 2019) and tutorials (Dozo, 2020) is to set the value to 100 because that is when another qualification (approval percentage) becomes active. This difference may be because other fields primarily use crowdsourcing for surveys rather than data annotation or human computation systems. It is unclear how representative these samples are. However, there are other sources that can provide information about conventions.

One source is Amazon itself. The Mechanical Turk web-interface provides six options: 50, 100, 500, 1,000, 5,000, 10,000. The MTurk blog has mentioned this qualification in four posts over the past eight years (Amazon Mechanical Turk, 2012, 2019, 2017, 2013). In three cases, the value was 5,000 and in the fourth it was 10,000.

Another source is forums and blogs. One pinned thread on the MTurk Crowd forum advises that "For your first 1000 HITs you may want to concentrate on approval milestones rather than $$$ ... most of the better-paying requesters require 1000/5000/10000+ approved HITs" ([jklmnop], 2016). This advice is repeated elsewhere on the forum and on Reddit ([WhereIsTheWork], 2019; [CaptainSlop], 2019; [Crazybritzombie], 2018). This is consistent with observations that 80% of tasks available to new users pay less than 10 cents (El Maarry et al., 2018). In one discussion between a worker and a requester, the worker recommended a threshold of 5,000 ([clickhappier], 2016). In the blog "Tips For Requesters On Mechanical Turk", one post recommends at least 5,000 if not 10,000 (Miele, 2012) while another recommends at least 1,000 (Miele, 2018). A web article by a Computer Vision researcher recommended 1,000 (Kumar, 2014). The CloudResearch blog mentions the threshold once, noting that a value of 10,000 maintains quality without significantly increasing the time to finish a set of HITs (Robinson, 2015).

Qualifications are also discussed by courses and tutorials. In the Crowdsourcing & Human Computation course at the University of Pennsylvania, a guest lecture on "The Best Practices of the Best Requesters" mentioned the approved HITs qualification and used 10,000 as an example (Milland, 2016). One guide recommends a cutoff of 5,000 (Carlson , née Feenstra).

Overall, we conclude that while practises vary, 5,000 or higher are commonly used as a qualification for tasks.

3.1 Impact on Workers

It is difficult to estimate how much time workers have to spend to achieve this qualification. Academic studies of time spent on HITs may be skewed by experienced workers, who have strategies for finding and completing tasks rapidly. Posts on Reddit mention taking anywhere from a month to a year to reach 5,000 approved HITs. The median of values reported across several Reddit threads was 2.25 months ([alisonlovepowell], 2015; [GnomeWaiter], 2013; [FrobozzYogurt], 2020; [Wat3rloo], 2016). Assuming 20 hours of work a week that is almost 200 hours of effort (140 seconds per task).

3.2 Potential Solutions

If this type of qualification undercuts our commitment to paying a fair wage, what are alternative ways to maintain quality? Options include:

1. Introduce screening questions that workers must complete correctly to proceed to the rest of the task, e.g., requiring 70%+ on three questions (Shvartzshnaid et al., 2019). This approach is prob-

lematic because it the workers who fail the screening are doing unpaid labour.

2. Address quality after collection by either dropping the lowest performing workers (e.g., the bottom 25% in Bansal et al., 2019), aggregating a larger number of responses per example, or including attention check questions and discarding workers who get them wrong. All of these incur a substantial cost to researchers.

3. Controlled crowdsourcing (Roit et al., 2020) uses an initial task that a broad set of workers can complete and then limits participation to the workers who did well on that task.[2] The cost of this solution depends on the percentage of workers who do well on the initial task.

4. Lower the threshold, reducing the required volume of earlier work. This reduces, but does not eliminate the qualification labour issue.

These methods can also be combined. Controlled crowdsourcing (method 3) with a very low Accepted HITs threshold (method 4) for the initial task would address the ethical concern we raise here while limiting the additional cost to the recruitment phase. Attention checks and aggregation (method 2) would then address natural variation in skill and attention during large-scale annotation.

4 Studying the Approved HITs Value

All of the options above have tradeoffs that will be task dependent and in practise some combination is most likely to be the best approach. The first three have been studied in prior work, but the impact of lowering the threshold has not. In this section, we consider the quality of work completed by workers grouped by how many tasks they have previously completed and what percentage were accepted.[3]

4.1 Tasks

Coreference Resolution This is an unusual task for crowdsourcing, with a novel user interface, shown in Figure 1. Workers were shown a 244 word document from the Ontonotes dataset (Hovy et al., 2006). We identified noun phrases using the Allen NLP parser (Gardner et al., 2018) and asked workers to identify when one of two specific entities was mentioned. This is not the complete coreference resolution task, but a useful subset. We refined the task over several rounds of trial annotation to ensure the instructions were clear and the interface was efficient. Workers were asked to check their answers if they tried to submit in less than 75 seconds. If they labeled 8 items in the first 19 words, they were reminded to only label the two entities specified. We estimated that the task would take 3 minutes and paid workers 60 cents ($12 / hour). Reviews on TurkerView (https://turkerview.com/) indicated that workers effective hourly rates were $7.88, $11.25, $12.93, and $14.59.

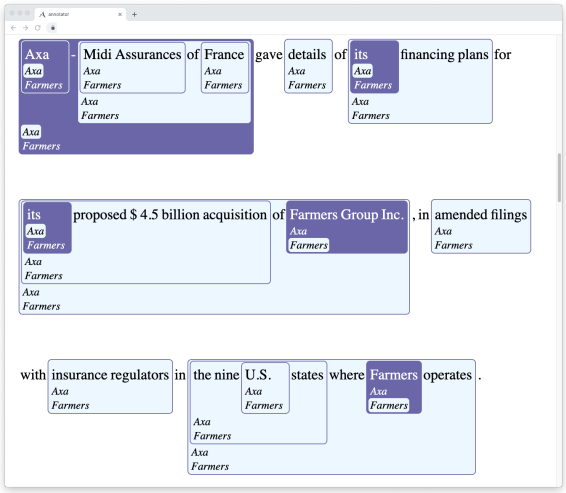

Figure 1: The user interface for coreference resolution (zoomed in). Spans are noun phrases automatically assigned by the Allen NLP syntactic parser (Gardner et al., 2018). The two entities being identified are the two most frequently mentioned entities in the text. Workers select a label by clicking on it.

We measure performance by comparing with the Ontonotes annotations. An F-score of 80% or above was considered acceptable, to allow for minor errors and points of confusion.

Sentiment Analysis This task is very intuitive and has been crowdsourced extensively in the past. We closely followed the set up used to annotate the Stanford Sentiment Treebank (Socher et al., 2013), with the same task instructions. Workers were shown ten examples whose true scores were evenly spread across 0 to 1. We estimated that the task would take 4 minutes and paid workers 80 cents ($12 / hour). Three reviews of the task on TurkerView indicated that workers hourly earnings were $22.15, $48.00, and $50.53, suggesting that workers were faster than anticipated.

[2] One potential drawback of this approach is that the filtering step may produce a biased sample of workers. That may be problematic for more subjective tasks, though with a large enough sample, responses could be weighted to make the results more representative.

[3] This was completed as part of a larger study approved by the Michigan IRB under study ID HUM00155689.

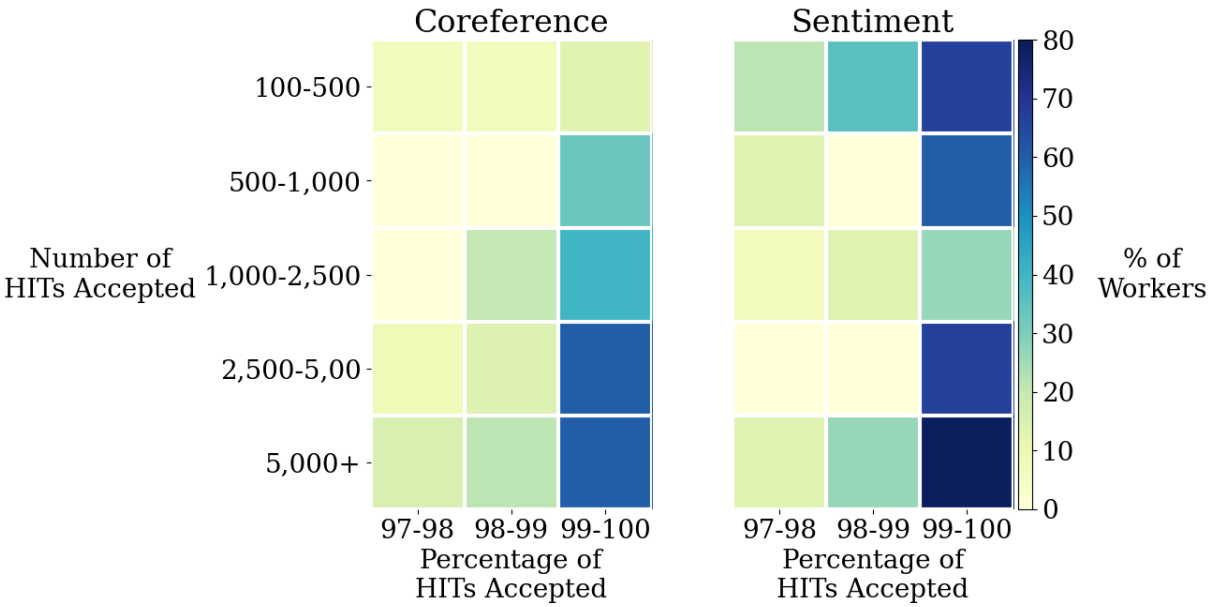

Figure 2: Results for all fifteen combinations of qualifications. Left (coreference): The percentage of workers scoring above 80 in each group. Right (sentiment): The percentage of workers whose average error was below 0.15 in each group. Each value is based on fifteen workers, except for sentiment there were fourteen for (98-99%, 500-1,000), and for coreference there were fourteen for (97-98%, 500-1,000), (97-98%, 1,000-2,500), (98-99%, 2,500-5,000), (98-99%, 5,000+), thirteen for (97-98%, 5,000+), and twelve for (97-98%, 2,500-5,000).

To evaluate, the labels are mapped to [0, 1] and compared with the STS values. An average value of below 0.15 was considered acceptable. This cutoff was chosen based on the scores achieved by two NLP students in our lab (0.11 and 0.09).

4.2 Recruitment

We considered 15 combinations of ranges for "Approved HITs" and "Percentage Approved", as shown by the axis labels in Figure 2. The ranges are based on the preset values provided by MTurk, with the addition of a boundary at 2,500 to provide slightly more detail in the shift between 1,000 and 5,000. Workers also had to be U.S.-based. We used Javascript-based checks to ensure each worker completed the task only once. 224 workers completed the sentiment task and 30 opened and returned it. 216 workers completed the coreference task and 657 opened and returned it. All but two conditions had 14 or 15 workers (the 97-98%, 5,000+ case for coreference had 13 and the 97-98%, 2,500-5,000 case for coreference had 12).

4.3 Results

The heatmap on the left of Figure 2 shows the percentage of workers scoring 80 or higher on the coreference resolution task. When the acceptance percentage is below 99, results are consistently poor, with fewer than 25% of workers scoring above 80. When the acceptance percentage is 99-100, groups with higher approved HITs have better scores. However, Figure 3 shows that more workers returned the HIT[4] in the groups with higher performance (see the last column of the rightmost plot), indicating that workers are self-selecting out.

This figure may be interpreted to suggest that a threshold of 2,500 is necessary. However, the distribution of workers is not uniform across these qualification groups. In a follow up experiment with constraints of 99-100% and 1,000+ using a relatively new requester account, 60 out of 92 workers scored 80 or above (65%), indicating that there are more workers in the higher approved HITs groups.

Figure 2 also shows results for the sentiment task. First, note that many more workers did well on the task. Comparing the left and right, the trend for percentage of HITs accepted is repeated, with consistently poor performance from workers with values below 99% (the left two columns). While the best result is the same in both cases (the bottom-right), the trend in the third column is somewhat different. Rather than a steady increase in performance as the approved HITs threshold increases, there is a U-shaped pattern. This shows that the pattern is somewhat task dependent.

[4] 'Returning' a task means a worker choose to stop working on it, receives no pay, but also receives no penalty in their profile for failing to complete the task.

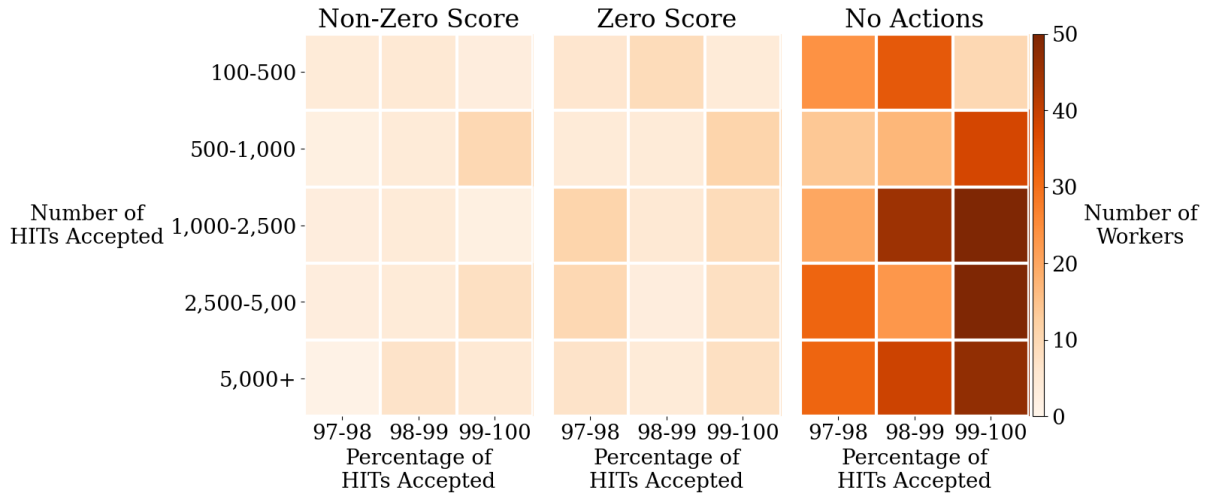

Figure 3: For coreference resolution, 657 workers opened and returned the HIT without completing it. These three heatmaps show the number of workers who: left partially correct annotations (Non-Zero Score), left entirely incorrect annotations (Zero Score), did not interact with the page (No Action). We do not include plots for sentiment analysis because only 30 workers opened and returned the HIT.

These results suggest that a lower qualification can be used without a substantial impact on work quality. In both tasks, the percentage HITs accepted qualification had a clear impact, with substantial decreases in quality from workers with a value below 99%. While that qualification does not directly force workers to do a substantial amount of work, it can be impacted by requesters who unfairly reject work. Our results also suggest that simply paying workers more will not lead to better work, as the sentiment analysis task paid considerably better and did not solve the issue.

5 Ethics and Impact Statement

This work involved consideration of several potential impacts. In terms of privacy, all data from workers is aggregated for the purpose of presenting results, and information from worker discussions were only sourced from publicly shared content. In terms of payment, we estimated the effort involved and aimed to pay workers $12 USD an hour. See the main text for worker reported values of hourly earnings on the two tasks. This was approved by the Michigan IRB under study ID HUM00155689. One potential harm of this work is that it may encourage higher values of the Percentage of HITs Accepted qualification, making workers more vulnerable to requesters who unfairly reject work.

6 Conclusion and Recommendations

This paper identifies the issue of Qualification Labour: the implied labour created by the qualifications we define. Based on a range of sources, we found that 5,000 approved tasks is one common threshold. That takes approximately two months to achieve and the tasks are poorly paid. We conducted a study of two tasks to understand how work quality correlates with these qualifications. We found that trends are task dependent, but lower thresholds can often be used.

We recommend either not using the ”HITs accepted” qualification, or running preliminary tests to identify the lowest suitable threshold for your task. This calibration is necessary because worker performance depends on many factors, including the task type, data (including which language), user interface, and instructions. One particularly promising method is to use controlled crowdsourcing (Roit et al., 2020) with a low threshold: run a short task with low or no qualifications to identify workers, then for the full task only allow those workers to participate. This reduces the burden on workers while maintaining high quality work.

Acknowledgements

We would like to thank Judy Kay, Ellen Stuart, Greg Durrett, attendees at the Conference on Human Computation and Crowdsourcing, and the ACL reviewers for helpful feedback. This aterial is based in part on work supported by DARPA (grant #D19AP00079), Bloomberg (Data Science Research Grant), and the Allen Institute for AI (Key Scientific Challenges Program).

References

[alisonlovepowell]. 2015. How long did it take you to hit 5000 completed hits? https://www.reddit.com/r/mturk/comments/3by1va/how_long_did_it_take_you_to_hit_5000_completed/. Accessed: 2020-08-12.

Amazon Mechanical Turk. 2012. Improving quality with qualifications – tips for api requesters. https://blog.mturk.com/improving-quality-with-qualifications-tips-for-api-requesters-87eff638f1d1. Accessed: 2020-08-12.

Amazon Mechanical Turk. 2013. Hit critique: Design tips for improving results. https://blog.mturk.com/hit-critique-design-tips-for-improving-results-a53eb8422081. Accessed: 2020-08-12.

Amazon Mechanical Turk. 2017. Tutorial: Understanding requirements and qualifications. https://blog.mturk.com/tutorial-understanding-requirements-and-qualifications-99a26069fba2. Accessed: 2020-08-12.

Amazon Mechanical Turk. 2019. Qualifications and worker task quality. https://blog.mturk.com/qualifications-and-worker-task-quality-best-practices-886f1f4e03fc. Accessed: 2020-08-12.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*.

Benjamin B. Bederson and Alexander J. Quinn. 2011. Web workers unite! addressing challenges of online laborers. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, pages 97–106.

[CaptainSlop]. 2019. Newbie that read faq's any tips to getting to 1000. https://www.reddit.com/r/mturk/comments/9bfv92/newbie_that_read_faqs_any_tips_to_getting_to_1000/e52rbs5/. Accessed: 2020-08-12.

[clickhappier]. 2016. Masters qualification info - everything you need to know. https://www.mturkcrowd.com/threads/masters-qualification-info-everything-you-need-to-know.1453/. Accessed: 2020-08-12.

[Crazybritzombie]. 2018. How to get to 5,000 approved hits? https://www.reddit.com/r/mturk/comments/90zzt5/how_to_get_to_5000_approved_hits/. Accessed: 2020-08-12.

Nerisa Dozo. 2020. Introduction to mturk and prolific.

Kinda El Maarry, Kristy Milland, and Wolf-Tilo Balke. 2018. A fair share of the work? the evolving ecosystem of crowd workers. In *Proceedings of the 10th ACM Conference on Web Science*, page 145–152.

Taylor Nicole Carlson (née Feenstra). 2014. Mechanical turk how to guide. http://pages.ucsd.edu/~tfeenstr/resources/mturkhowto.pdf. Accessed: 2020-08-12.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

[FrobozzYogurt]. 2020. Just hit 100k! https://www.reddit.com/r/mturk/comments/i3nvx4/just_hit_100k/. Accessed: 2020-08-12.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

[GnomeWaiter]. 2013. Over $1100 and 5000+ approvals in my first month of turking, and so can you! https://www.reddit.com/r/mturk/comments/1tjge3/over_1100_and_5000_approvals_in_my_first_month_of/. Accessed: 2020-08-12.

Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 321–329.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–14.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 611–620.

Lilly C. Irani and M. Six Silberman. 2016. Stories we tell about labor: Turkopticon and the trouble with "design". In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4573–4586.

[jklmnop]. 2016. Your first 1000 hits. https://www.mturkcrowd.com/threads/your-first-1000-hits.23/. Accessed: 2020-08-12.

Neeraj Kumar. 2014. Effective use of amazon mechanical turk (mturk). https://neerajkumar.org/writings/mturk/. Accessed: 2020-08-12.

Pietro Michelucci Libuše Hannah Vepřek, Patricia Seymour. 2020. Human computation requires and enables a new approach to ethical review. In *Proceedings of the NeurIPS Crowd Science Workshop*.

David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 224–235.

Joe Miele. 2012. Tips for academic requesters on mturk. http://turkrequesters.blogspot.com/2012/09/tips-for-academic-requesters-on-mturk.html. Accessed: 2020-08-12.

Joe Miele. 2018. The bot problem on mturk. http://turkrequesters.blogspot.com/2018/08/the-bot-problem-on-mturk.html. Accessed: 2020-08-12.

Kristy Milland. 2016. The best practices of the best requesters. http://crowdsourcing-class.org/slides/best-practices-of-best-requesters.pdf. Accessed: 2020-08-12.

Tanushree Mitra, C.J. Hutto, and Eric Gilbert. 2015. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, page 1345–1354.

Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on paid crowdsourcing platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–14.

Jonathan Robinson. 2015. Maximizing hit participation. https://www.cloudresearch.com/resources/blog/maximizing-hit-participation/. Accessed: 2020-08-12.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769.

Yan Shvartzshnaid, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. 2019. Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Colin Vandenhof. 2019. A hybrid approach to identifying unknown unknowns of predictive models. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*.

[Wat3rloo]. 2016. 5000 approved hits!!?? https://www.reddit.com/r/mturk/comments/4kd1co/5000_approved_hits/. Accessed: 2020-08-12.

[WhereIsTheWork]. 2019. How important are qualifications for getting more surveys? https://www.mturkcrowd.com/threads/how-important-are-qualifications-for-getting-more-surveys.4521/. Accessed: 2020-08-12.

Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*.

Jacob Young and Kristie M. Young. 2019. Don't get lost in the crowd: Best practices for using amazon's mechanical turk in behavioral research. *Journal of the Midwest Association for Information Systems (JMWAIS)*.