Qualification Labour: A Fair Wage Isn't Enough if Workers Need to Do 5,000 Low Paid Tasks to Qualify for Your Task

Jonathan K. Kummerfeld
Computer Science & Engineering
University of Michigan
Ann Arbor, Michigan, 48109
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 (e.g. for Natural Language Processing) suggests using qualifications such as a minimum number of previously completed tasks. If most requesters who pay fairly use this kind of minimum qualification, then workers may be forced to complete a substantial amount of poorly paid work for other requesters before they can earn a fair wage. This paper (1) explores current conventions for the threshold, (2) discusses possible alternatives, and (3) presents a study of correlation between approved work and work quality.

The Problem

Workers using Amazon Mechanical Turk earn a median wage of $2.54 an hour while completing tasks (Hara et al. 2018), far below U.S.-state minimum wages of $7.25 to $15. Many researchers try to pay workers a higher wage, carefully estimating the time spent on a task and giving workers 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, Hutto, and Gilbert 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 this restriction is widely used by high paying requesters it means we are requiring workers to do a substantial amount of low paid "Qualification Labour": work to achieve the qualifications necessary for high paying tasks.

It is difficult to estimate how much time workers have to spend to achieve these qualifications. 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 four 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).

Conventions for the Approved HITs Threshold

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, Hugh, and Bernstein 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 these fields primarily use crowdsourcing for surveys rather than data annotation or human computation systems. It is unclear how representative the samples listed above are. However, there are other sources that can provide information about conventions.

One source is Amazon itself. The Mechanical Turk web-interface provides six threshold 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 recommended value is 5,000 and in the fourth it is 10,000.

Another source is forums and blogs. One pinned/sticky 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). In 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). In the CloudResearch blog, the threshold is mentioned once, noting that a value of 10,000 maintains quality without significantly increasing the time to finish a set of HITs (Robinson 2015).

Finally, qualifications are 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) 2014).

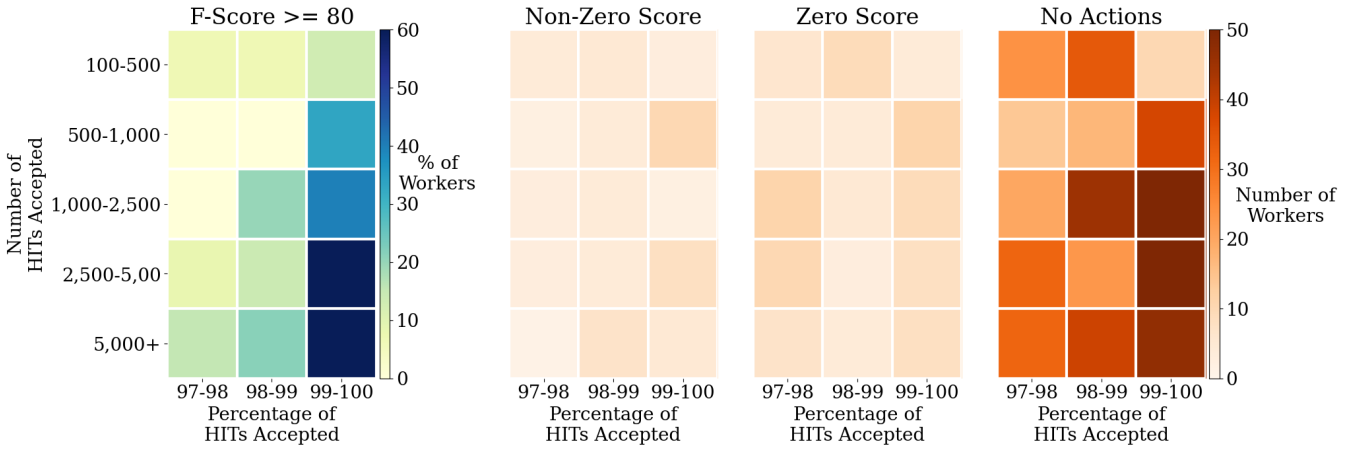

Figure 1: Results for fifteen combinations of qualifications. Left (workers who finished): The percentage of workers scoring above 80 in each group. Right (workers who returned the HIT): # who left partially correct annotations (Non-Zero Score), # who left entirely incorrect annotations (Zero Score), # who did not interact with the page (No Action).

Potential Solutions

If this type of qualification undercuts our commitment to paying a fair wage, what are alternative ways to maintain quality? One option is to introduce screening questions that workers must complete with a certain accuracy to proceed to the rest of the task, e.g., requiring 70%+ on three questions with known answers (Shvartzshnaid et al. 2019). The problem with this approach is that it involves unpaid labour from workers who fail to pass the screening. Another option is to simply accept that there will be lower quality workers and drop the lowest performing ones, e.g., the bottom 25% (Bansal et al. 2019). This incurs a substantial cost to the researchers, as do other approaches that involve aggregation of responses or attention checks. Finally, there is the option to have an initial task that a broad set of workers can complete and then limit participation to the workers who did well on that task. The cost of this solution depends on the percentage of workers who do well on the initial task. One drawback of this solution is that the filtering step may produce a biased sample of workers, though with a large enough sample that could be corrected for by weighting if needed.

Calibrating the Approved HITs Threshold

In the approaches above, the approved HITs qualification is either removed entirely or the threshold is set to a lower value. This section presents a study of the quality of work performed by workers in different ranges for the threshold.[1]

Task: Workers were shown a 244 word document and asked to identify when one of two specific entities was mentioned. 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. The task was estimated to take 3 minutes and paid workers 60 cents ($12 / hour). Four reviews of the task on TurkerView (https://turkerview.com/) indicated that workers earned $7.88, $11.25, $12.93, and $14.59.

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

Recruitment: We considered 15 combinations of ranges for "Approved HITs" and "Percentage Approved", as shown by the axis labels in Figure 1. The task was hosted externally, with Javascript-based checks to ensure each worker completed the task only once. Workers also had to be U.S.-based. 216 workers completed the task and 657 opened and returned it. Most groups had 15 who finished, with 14 in four cases (97-98% with 500-1,000 and 1,000-2,500 approved, and 98-99% with 2,500-5,000 and 5,000+), 13 in one (97-98%, 5,000+) and 12 in one (97-98%, 2,500-5,000).

Results: For this task, an F-Score of 80 is a reasonable threshold for having carefully read the instructions and attempted the task. Figure 1 shows the percentage of workers scoring 80 or higher (leftmost plot) and information about the behaviour of workers who returned the HIT. When the acceptance percentage is below 99, results are consistently poor, with fewer than 25% of workers scoring above 80. When acceptance percentage is 99-100, groups with higher approved HITs have better scores. However, the number of workers returning the HIT is also higher in these cases (see the last column of the rightmost plot), indicating that workers are self-selecting out. Finally, in a follow up experiment with constraints of 99-100% and 1,000+ Approved HITs and 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.

Conclusion

This paper identifies hidden qualification labour and explores ways to reduce it. The common practise of requiring 5,000 approved tasks means workers need to complete approximately two months of work at extremely low rates. This work also considers the impact of changing the threshold on the quality of work for an example Natural Language Processing task. Accuracy correlates with the number of tasks completed, though percentage accepted is more critical, and there is a major shift at the 1,000 approved tasks mark. Shifting to 1,000, as some researchers already have, would substantially reduce qualification labour.

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.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 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*.

[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.

Carlson (née Feenstra), T. N. 2014. Mechanical turk how to guide. http://pages.ucsd.edu/~tfeenstr/resources/mturkhowto.pdf. 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.

Dozo, N. 2020. Introduction to mturk and prolific.

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

[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.

Hara, K.; Adams, A.; Milland, K.; Savage, S.; Callison-Burch, C.; and Bigham, J. P. 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*, 1–14.

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

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

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

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

Mitra, T.; Hutto, C.; and Gilbert, E. 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*, 1345–1354.

Oppenlaender, J.; Milland, K.; Visuri, A.; Ipeirotis, P.; and Hosio, S. 2020. Creativity on paid crowdsourcing platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

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

Shvartzshnaid, Y.; Apthorpe, N.; Feamster, N.; and Nissenbaum, H. 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*.

Vandenhof, C. 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.

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

Young, J., and Young, K. M. 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)*.