Task Difficulty on Worker Engagement on Mechanical Turk

John Oliver
Cornell University

Introduction

Crowdsourcing networks such as Mechanical Turk and TaskCN, provide the power to solve many problems currently intractable to a purely computational strategy. By highly parallelizing a task among thousands of human workers, these problems can be solved quick and at a reasonably expense.

However, as is often the case when bringing humans into the equation, keeping the quality of work high can be difficult: workers have differing motivations and skill levels, often resulting in highly variable levels of quality for even simple tasks.

Previous literature such as (Ipeirotis et al. 2010,) summarizes common approaches to maintaining worker quality. The most common approach is to apply redundancy to the problem: give the task to a number of workers and take the majority opinion. This method can successfully identify the correct answer given enough redundancy, however, this method multiplies total costs and quickly becomes infeasible for any large experiment.

Further work has been done to estimate the overall quality of a worker, generating a "confusion matrix" for each worker with individual estimated error rates for a given task (Dawid and Skene 1979,), and work from (Ipeirotis et al. 2010,) attempts to correct workers for bias (where workers may be putting in effort, but don't properly understand the details of the task).

Pre-filtering methods have also been implemented (namely by the crowd sourcing platforms themselves) offering potential workers a quick survey or sample tasks to attempt to filter out those who perform badly for any number of reasons. This is not a perfect solution, though, and potentially reduces your potential pool of workers to a limited group of high-performers.

These techniques are effective within their domains (the often tested example is object labeling), but traditional quality management methods are difficult to implement when working with tasks with subjective validation. For example, a common task on Mechanical Turk has workers write summaries of online articles. Here automated validation is difficult as there is no *ground truth* with which to compare against. Additionally, as there is no one exact answer, group majority voting can't be applied. Peer evaluation methods using separate validation tasks fed back to workers can be used to check subjective task quality, however these methods increase overall task complexity as well as monetary overhead.

Flow and Worker Engagement

An area often overlooked by the current literature is how the engagement of a worker affects the overall quality of their output. From previous work, it is known that 41% of turkers claim to do these tasks for *enjoyment* over strictly monetary gain (Paolacci et al. 2010,). Further, (Eickhoff et al. 2012,) finds that workers perform better on attractive tasks (those that are more 'game-like'). These two results suggest that workers highly value engaging and enjoyable tasks. From this, the question then arises:

How engaged are mechanical turk workers, and how can we keep them more engaged?

To begin to answer this, it is useful to look to the field of psychology and specifically flow research. This area focuses on human experience and development, looking closely at what drives motivation, and keeps people engaged. A highly cited reference in this field is (Nakamura and Csikszentmihalyi 2002,) which defines a state of being in "flow" as 'the state of extreme focus attained when a person is deeply invested in a job'. Examples of being in flow might be an artist working on a new painting or a player engaged with a video game. While in this state, people have the following traits:

- Intense and focused concentration on what one is doing at the present moment.

- Experience the activity as intrinsically rewarding.

These are both qualities that are highly coveted in a strong mechanical turk worker.

Even further, Nakamura details specific conditions in which this flow occurs, which include:

- Perceived challenges, or opportunities for action, that stretch existing skills; a sense that one is engaging challenges at a level appropriate to one's capacities

- Clear proximal goals and immediate feedback about the progress that is being made

In this paper we attempt to apply these considerations to Mechanical Turk task design. The second condition is largely satisfied by the instant positive feedback of mechanical turk, so we focus on the first. We look at how the difficulty of tasks, and the juxtaposition of task challenges can affect worker effort and overall accuracy.

Experimental Task Design

We first designed a Mechanical Turk task to measure the effort a worker puts into a specific task. We explored a few avenues (hidden object games, reflex games) but eventually settled on a simple "Count the numbers" game. See figure 1. Workers were shown a set of anywhere from 20 to 60 numbers depending on the difficulty of the challenge, and given a *target number* and a *guess*. They were then asked to count how many times the *target number* appeared in the overall set and label the task 'Correct' if the number matched the *guess*, otherwise 'Wrong'. A simple python script was written to generate these tasks and write them into the Mechanical Turk data format.

We chose this type of task because it was simple to generate, easy to explain to users, and as a binary labeling task, did not require the complex setup that comes from other effort measuring tasks (as in (Horton and Chilton 2010,))

Additionally, the task has the following traits:

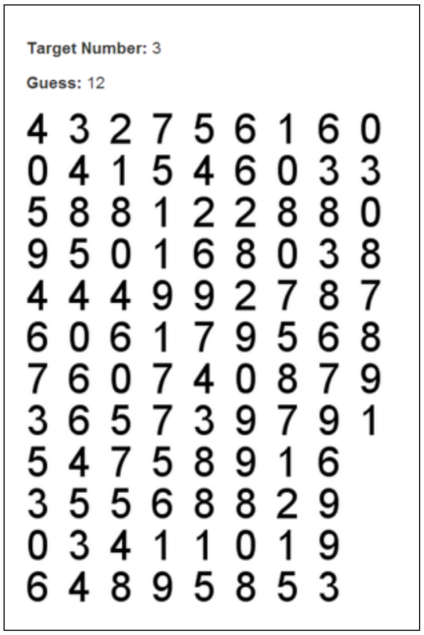
- **Validatable** - For analysis, we would like to be able to see how varying effort affects overall accuracy. The task is quick to validate against against the ground truth. Additionally, turkers are unlikely to achieve an accuracy better than guessing on a single task unless full effort is exerted to count each number.

- **Scalable Difficulty** - The difficulty of the task is simple to control without much margin of error. Adding more numbers increases the time workers must sustain effort and memory to succeed.

- **Effort Based Difficulty** - A harder task requires more effort to complete (the worker must analyze more numbers). Also, a user's capacity to complete a task won't increase much with repeated work, unlike a more complex puzzle.

Experiment Design

We ran three Mechanical Turk category tagging task, splitting participants into three groups. The first group (41 tasks) received only easy tasks; the second (39 tasks) received only hard tasks; the last group (107 tasks) received a randomized mix of both easy and hard tasks.


Target Number: 3

Guess: 12

```
4 3 2 7 5 6 1 6 0
0 4 1 5 4 6 0 3 3
5 8 8 1 2 2 8 8 0
9 5 0 1 6 8 0 3 8
4 4 4 9 9 2 7 8 7
6 0 6 1 7 9 5 6 8
7 6 0 7 4 0 8 7 9
3 6 5 7 3 9 7 9 1
5 4 7 5 8 9 1 6
3 5 5 6 8 8 2 9
0 3 4 1 1 0 1 9
6 4 8 9 5 8 5 3
```

Figure 1. A sample 'hard' task. Here users were asked to count how many '3's there were and compare that to a guess of 12

After receiving results, we looked at the average response time to verify that the difficulty of the task correlated with effort. A group with medium difficulty was used in this measurement, although it was not included in the final experiment.

From figure 2 we can see that harder difficulties required a longer amount time to complete and thus more sustained effort, which is consistent with our initial task design.

	Easy	Medium	Hard
Avg. Time (s)	18	35	54

Figure 2. Average time taken in seconds for tasks of varying difficulty.

Results

We looked at the overall accuracy between the three groups of participants. The results can be found in figure 3.

When presented with only easy tasks, workers performed phenomenally, with an overall accuracy of 95%. Conversely, when presented with only hard tasks, the accuracy dropped significantly down to 56%, or no better

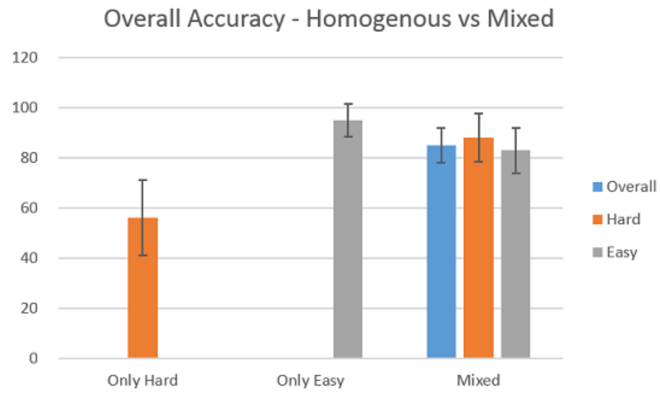


Figure 3. Task accuracy for each group. The 'mixed difficulties' group is further split into individual accuracies for 'hard' and 'easy' within that task. Error calculated at 95% significance.

than guessing. Intriguingly, when both easy and hard challenges were mixed, the individual accuracy of the hard challenges was raised up to 88% and the individual accuracy of easy challenges maintained a similar level.

More exploration and evaluation is necessary to pinpoint the exact cause of these results, but an intuitive reasoning might be the following:

When faced with only hard challenges, workers seemed to ignore the challenges or not put in the required effort to accurately guess the correct response. A possible explanation for this may be an excessive cognitive load. Additionally, this scenario may have enticed people to act more rationally towards the monetary incentive, as payments were not tied to output quality.

With only easier tasks, the tasks are easy enough to skim quickly and answer correctly. Given that an easy question only took 18 seconds, there was little benefit to guessing randomly over simply performing the task properly.

When both types were mixed, a reasonable explanation for the better performance on the more difficult tasks is that workers were highly motivated by the easier tasks and the quick money made, and that motivation carried across to the harder tasks. Additionally, the increased variety may keep workers from becoming bored with the task (even if they find it easy enough to complete).

A different explanation from flow research might be that having tasks of multiple difficulties keeps the worker within the *flow channel*, which is a 'goldilocks zone' within the graph of player skill and task difficulty. While a number of hard tasks overwhelms the worker into anxiety, interleaving easy tasks balances out the overall perceived difficulty. Flow theory would also predict many easy tasks to be too boring (and thus have low effort), however the presence of external incentives within the crowd-source framework likely overcame boredom in this case.

Task Design

These results suggest there to be value in varying task difficulty, and for those problems in which the difficulty is known without the solution, varying the difficulty of tasks on a per-worker basis should keep workers more engaged, and allow for higher quality output.

Additionally, for those challenges in which there are a few significantly more difficult tasks, a designer should interleave easier tasks with the harder ones to keep worker motivations strong.

Future Work

While the results found here are intriguing, further work should be done to validate these results in the general case. For instance, it would be useful to run a similar experiment with multiple types of tasks, and with a broader selection of workers. With a larger experiment, the categories of difficulty could be broken down more granularly, and a broader correlation could be assessed.

A second consideration is with different types of workers. It is fully possible that some users are simply more engaged by challenging tasks than others. A useful path would be to design an experiment to analyze whether there exists certain groups of users who are motivated more or less by task difficulty.

Lastly, previous research (Andersen et al. 2013,) indicates that in addition to varying the difficulty, the order of tasks can have a significant affect on the engagement of a student. For instance, when outlining learning patterns for introductory algebra classes, the authors found that giving an easy task immediately followed by the most difficult task maximized user engagement throughout the rest of the educational process. This was not a possible avenue given the smaller scale of this experiment, but would be an obvious next step to look at exactly how users are motivated.

Overall this study should be taken as a good starting point for further investigation into the affects of task difficulty on overall worker quality. Workers on mechanical turk are not perfect rational machines, and by understanding methods for motivating workers, high quality results can be attained at the benefit of both requester and worker.

References

ANDERSEN, E., GULWANI, S., AND POPOVIC, Z. 2013. A trace-based framework for analyzing and synthesizing educational progressions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 773–782.

DAWID, A. P., AND SKENE, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 20–28.

EICKHOFF, C., HARRIS, C. G., DE VRIES, A. P., AND SRINIVASAN, P. 2012. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, ACM, 871–880.

HORTON, J. J., AND CHILTON, L. B. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, ACM, 209–218.

IPEIROTIS, P. G., PROVOST, F., AND WANG, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, ACM, 64–67.

NAKAMURA, J., AND CSIKSZENTMIHALYI, M. 2002. The concept of flow. *Handbook of positive psychology*, 89–105.

PAOLACCI, G., CHANDLER, J., AND IPEIROTIS, P. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making 5*, 5, 411–419.