Financial Incentives and the Performance of Crowds Replication

Kevin Zhai, Samay Dhawan, Ella Polo, Victor Yoon

I. Literature Review

For our literature review, we decided to look at three papers: Demographics of Mechanical Turk, Labeling Images with a Computer Game, and Financial Incentives and the Performance of Crowds. We chose the last of these to replicate.

Demographics of Mechanical Turk

Since Amazon changed their compensation policy to allow workers in India to be paid rupees, the demographics of Turkers has changed dramatically. The researcher, Panos Ipeirotis from the Stern school of business in NYU, conducted a demographics survey by creating a HIT on the Mechanical Turk that simply asked the users about the reasons they used the platform and their demographic data. These data points ranged from gender to country of origin, and from household income to educational background. Surveys were collected from 1000 Turkers who were paid $0.10 to answer the questions, and the results were that 46% percent were American and 34% were Indian. Based on this clear dichotomy of workers, Ipeirotis decided to focus his demographic analysis on the two ethnic pools of workers.

Overall, the paper found some very interesting results. The majority of American turkers were female, while the majority of Indian turkers were male. Additionally, 55% of Indian turkers made less than $10,000 a year, as opposed to the 7% of Americans. This reveals a striking difference in the reason for each group to work on the platform. Most Indian workers relied on MTurk to supplement their income, while many Americans simply did it for sport. Overall, the paper has some good findings, but perhaps the sample size and response bias would skew some of the results that were obtained.

ii. Labeling Images with a Computer Game

The paper deals with the problem of accurately labeling images on the web. The authors, von Ahn and Dabbish intended to create a “game that is fun and could be used to create valuable output”. The game, called the “ESP game”, pairs individuals randomly and provide labels for any given image currently presented to them. The goal of each individual is to guess what their partner is typing for each image. In order to move on to the next image, both partners have to type the same string in reference to the image currently on the screen. Both partners have 2.5 minutes, and attempt to agree on as many images as they can within the time frame. Essentially, by making the goal of the game to guess the partner's string as opposed to labeling images alone, partners will try harder to come up with more generic, common terms that describe the image they're looking at - this forces players to try and think like each other. As the paper points out, the common string the players agree on tends to be a good label for the image.

The game also uses taboo words, or words that players are not allowed to enter in as guesses in successive rounds - the purpose of these words are to refine the quality of the words as the rounds go on. For example, if in the first round the players agree on more generic words to describe the words, and those words become “taboo” words, players might use more refined terms to guess on, and agree on those guesses. As a result, more specific words are agreed upon, and the threshold of word quality is increased.

In terms of the success the game has had both in usage and in successful labels associated with images. From August 9 2003 to December 10, in a period of a month, they had 13,630 people play the game, generating 1,271,451 labels for 293,760 images. Some players spent over 50 hours playing the game - clearly, they found merit in both the fun and productivity value of the game. In essence, the point of the game was to show that, rather than developing a more complicated algorithm, a large-scale problem can be solved with a method that fully utilizes the people active on and playing on the web.

iii. Financial Incentives and the Performance of Crowds

The authors of this paper analyzed how the compensation offered to workers on Mechanical Turk affected their performance on a task. In the first experiment, they set up a task on AMT that involved sorting traffic images, which was a task that allowed measurement of both accuracy and output. Participants completed a survey and then were automatically sorted into three pay groups (low, medium, high), and three difficulty groups (easy, medium, hard). There were also some groups that received bonuses. The first finding was that higher compensation and higher difficulty increased and decreased the number of completed tasks, respectively. They used a carefully defined measure of accuracy to conclude that the pay per task had _no effect_ on the accuracy of the result. They concluded that this was due to an “anchoring effect”, where the first compensation level offered to the workers framed their idea of the value of the task. Because of this effect, every group of workers felt they deserved higher compensation than they received, so no group was motivated to perform better on the task.

The experimenters also conducted a second study involving a word puzzle task. The word puzzle task was designed to control for possible issues with the traffic sorting task, like that some variations were too easy or that increased effort could not lead to higher accuracy. In this task, participants had to find words from a given list in a word puzzle. Since not all the words in the list were actually in the puzzle, participants had to work harder to find all the hidden words -- more effort could lead to higher accuracy. They introduced two payment schemes, each with three difficulty levels. In one scheme, participants were paid per word found, and in the other they were paid by puzzles completed. As in the previous experiment, they found that accuracy did not increase with an increase in pay. This time, probably because of intrinsic motivation, the quantity of work also did not increase in response to higher pay.

II. In-Depth Lit Review

Because we chose to replicate the first experiment in the Financial Incentives paper, we will focus on that experiment here and delve into the methods used by the authors to gather the data. The task in this experiment was, as stated above, traffic image ordering. These were photos taken at two second intervals and randomly shuffled, and the participants in the study had to place them in chronological order. This allowed for objective measurement of both the output and the accuracy of the workers.

The authors recruited crowdworkers using Amazon's Mechanical Turk platform. Participants were offered a base pay of $.10 to complete the introductory survey and training, and then had the option to complete up to 99 of the image ordering tasks. Once the crowdworkers completed the training, they were randomly assigned to a difficulty level (sorting two, three, or four images) and to a pay level (nothing, $.01, $.05, or $.10 per task). The experimenters most likely achieved this experimental design by linking from AMT to their own website that controlled the task, which was a drag-and-drop activity not supported by AMT, as well as the randomization into different experimental groups. The users were shown an initial demographic survey, the succession of drag-and-drop sorting tasks, and a final feedback survey at the end. This final survey was useful for understanding how satisfied people were with the level of pay they received.

When the experiment was completed, 611 crowdworkers had worked on the task, and they had sorted 36,425 images. They found that crowdworkers completed more tasks when they were paid more and completed less tasks when there were more images per set to sort. To measure the accuracy of participants, the experimenters measured both the proportion of completely accurate image sets as well as the squared difference between the number correct and total number of images in a set. In the paper, they reported the percentage measure of accuracy across difficulty levels and pay levels in a line graph, and found that there was no statistically significant difference in accuracy between the pay levels. Of course, the accuracy varied by the difficulty of the task (measured by number of images), but the difficulty did not affect the change in accuracy across pay levels. To confirm this result, the authors constructed a hierarchical linear model, which modeled the probability that each task was sorted correctly, taking into account both variability among users as well as variability among the tasks. They presented the coefficients of this regression, and the coefficients for each of the pay levels included zero, indicating that there was no reliable effect of pay on accuracy.

In the second study, which we will cover in less depth since we did not choose it, the experimenters changed the task to a word puzzle, in which participants were tasked with finding words in a grid of letters. Here the goal was to be able to better measure increased effort, as participants could find only some of the hidden words before moving on to the next set. They did not know how many of the listed words were hidden in the puzzle, so the number of words found was a more accurate measure of increased effort. In this experimental design, there were two different pay schemes: a quota pay scheme, where crowdworkers were paid based on the number of puzzles completed, and a "piece rate" scheme, where workers were paid based on the number of words they located in the puzzles. As in the first experiment, there were four pay levels: nothing, $.01, $.05, and $.10.

In this experiment there were 320 participants who in total completed 2736 word puzzles. They found some bias in the gender distribution for this experiment, with significantly more women than men. This suggested that women on AMT may have been more intrinsically motivated to complete the word puzzles, and therefore that intrinsic motivation may have affected both amount of tasks completed and the accuracy (regardless of pay level). This may have explained the finding that the pay scheme did not affect the output of the workers. They measured accuracy as the proportion of words found in each puzzle and ran the hierarchical linear model again. The findings were the same as in the first experiment -- namely, that the compensation scheme had no effect on the accuracy of the workers. They summarized these findings, as in part one, using a line graph of pay scheme vs. accuracy and with a chart containing coefficients for the hierarchical linear model.

Experiment Selection

Of the two studies conducted in the original paper, we decided to replicate the image ordering study. Our primary reasoning for choosing the first study was that we were more interested in the effects of financial incentives on the quantity and quality of work performed as a whole, rather than the nuance of how different pay schemes affect performance. Furthermore, we believed that the methodology of the image ordering study better suited our team's skill set. The format of presenting image links for the image ordering HITs was fairly simple to replicate using tools from the "Become a Requester" assignment. On the other hand, replicating the actual word puzzle component of the second study, as well as the varying pay schemes, would have proved particularly challenging given our relative inexperience with CrowdFlower and Amazon's HIT platforms.

Our Experimental Design

We want to replicate the structure used by the financial incentives experiment. We used the Crowdflower platform since there were already funds loaded in our accounts, and we thought it would be interesting to see whether choice of platform affected the outcome of the experiment. We decided to replicate the first experiment in the paper, where the experimenters had contributors order traffic images under different tiers of compensation. Because we do not have access to unordered traffic images, we decided to use images of clocks instead. In our design, crowdworkers were shown image sets of analog clocks and told to order them chronologically, assuming time in the AM. They ordered the images by answering multiple choice questions asking which image should be ordered first, second, and third. Because difficulty level had no effect on the relationship between pay and accuracy, we decided to eliminate these difficulty levels in our experiment and use the average from the traffic image experiment: three images. We used the same pay levels as the paper: low pay ($.01), medium pay ($.05), and high pay ($.10). Refer to the methodology for more information on how we collected the images of the clocks, created random pairings of 3 clocks and exported them to a .csv file, created the HIT, and collected the data from the crowd workers.

III. Methodology

i. Clock Face Collection

In our experimental design, we first needed to find a way to present a unique set of image data on CrowdFlower for the crowd workers to have access to in order to complete the task. We started with a list of 1000 unique clock urls that we obtained from an image database, image-net.org. Once we obtained the list of unique urls, we needed to tackle the most time consuming portion of the task, sifting through the urls to both note the time on the clock, and note if any of the urls were either junk or unreadable. The final product we had, as a result of this process, was **“times-aggregated.txt”**, a file with decimal values corresponding to the times on the clock (10.13 = 10:13, 0.20=12:20), followed by the urls that lead to the image of the clock. Our reason behind noting the time was to find a way to automatically order the pairings of clocks we eventually created, so that we could eventually cross-check the accuracy of the crowdworker output with the true clock orderings.

Our next step was to find a way to create random pairings between the urls we had obtained, to provide as input to CrowdFlower. To automate this process, we wrote **“create_task.py”**, a script that would create random combinations of 3 clocks, and outputted those combinations into a file we called **“clock_combos_data.csv”**. This was the data that we provided to crowdflower.

ii. HIT Interface

Once we finalized our .csv file, **“clock_combos_data.csv”**, we created our HIT interface. We wanted to design our HIT so that it looked similar to the image ordering HIT within the financial incentives paper; however, we lacked the luxury of creating an external tool from which users could work. In spite of this, we tried to be as “by the book as possible”. This was the design of the HIT:

Table X: Image Categorization

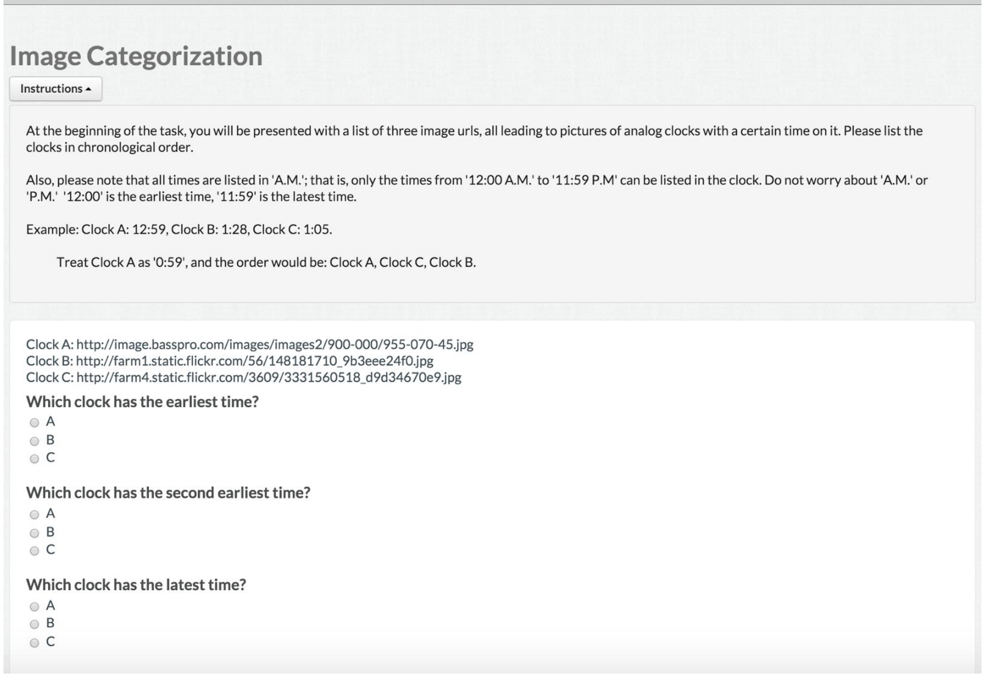


Image Categorization

At the beginning of the task, you will be presented with a list of three image urls, all leading to pictures of analog clocks with a certain time on it. Please list the clocks in chronological order.

Also, please note that all times are listed in 'A.M.'; that is, only the times from '12:00 A.M.' to '11:59 P.M' can be listed in the clock. Do not worry about 'A.M.' or 'P.M.' '12:00' is the earliest time, '11:59' is the latest time.

Example: Clock A: 12:59, Clock B: 1:28, Clock C: 1:05.

Treat Clock A as '0:59', and the order would be: Clock A, Clock C, Clock B.

Clock A: http://image.basspro.com/images/images2/900-000/955-070-45.jpg
Clock B: http://farm1.static.flickr.com/56/148181710_9b3eee24f0.jpg
Clock C: http://farm4.static.flickr.com/3609/3331560518_d9d34670e9.jpg

Which clock has the earliest time?

○ A
○ B
○ C

Which clock has the second earliest time?

○ A
○ B
○ C

Which clock has the latest time?

○ A
○ B
○ C

We presented the worker with instructions for the task at the top of the HIT, and then presented the user with three URLS, and asked them to label the clocks as “earliest”, “second-earliest”, and “latest”. Unfortunately, we could not find a way to embed the images within the HIT using crowdflower, but that is something that we would have liked to do to increase the quality of the HIT design. However, we proceeded with this and launched the HIT, paying workers variable amounts of $0.05, $0.25, and $0.50 per page, with 5 rows (essentially 5 tasks labeling clocks) per page. This came out to $0.01, $0.05, and $0.10 per labeling task, which mirrored the image ordering experimental setup of the paper.

iii. Accuracy Validation

After running the HIT at the variable amounts of pay on crowdworkers, we obtained a .csv file, with each of the pairs of urls and the corresponding labels (“a” for earliest, “b” for second-earliest, and “c” for latest). Within “create_task.py”, we wrote a “validator” script, that would order the pairs of urls (using their associated times stored in a dictionary) and check the automatic ordering with the orderings we obtained from crowdflower. This, along with the script we used to create the orderings, and all the data sources, From this, we found an accuracy of ~66% from the $0.01 orderings, ~71% from the $0.05 orderings, and ~69% from the $0.10 orderings. Refer to the “Results” section below for a more in-depth analysis on the results we obtained from the CrowdFlower experiment.

iv. Deviations from the original paper

Unfortunately, whether for lack of access to more sophisticated tools or because of constraints on our abilities to execute a professional crowdsourcing experiment, we deviated from the original design of the image ordering experiment in several ways:

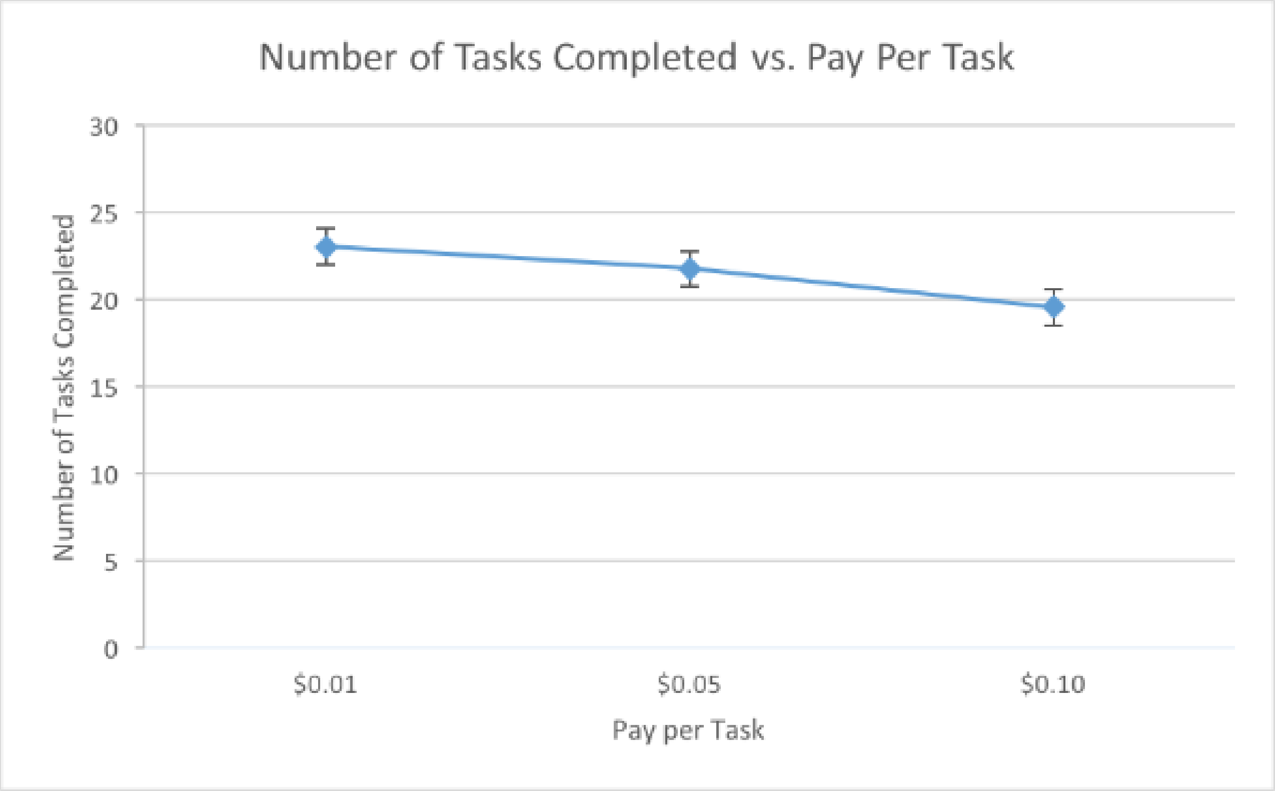
1. We decided to proceed with a set of random orderings of clocks, as we wanted to create our own, image set to utilize on crowdflower, unable to gain access to the original car image sets.
2. We posted the tasks within the CrowdFlower platform, as opposed to separately, hosted on our own web application. The original experiment seemed to redirect to a link
3. Within our HIT, we asked the workers to order the clocks in the form of multiple choice questions, as opposed to the drag-and-drop method done originally.
4. The original experiment had difficulty levels (3, 4, and 5 image orderings per set), whereas we withheld from creating HITS with different difficulty levels (mainly because each new HIT was draining a significant portion of our funds).
5. As opposed to embedding the images within the HIT, we embedded urls that the workers had to click on to see the clocks (probably added to the difficulty of the task and contributed to accuracy in some way)
6. We used CrowdFlower as opposed to Mechanical Turk.
7. Our users didn't have to go through a training set or take a survey before they began.

V. Crowdsourcing Platform

We opted to use CrowdFlower as our crowdsourcing platform for the image ordering HITs because of our experience using the platform in the last assignment. In accordance with the original paper, we paid workers in each experimental group $0.01, $0.05, or $0.10, respectively. For each HIT, we presented workers links to three different images of clocks, labelled 'A', 'B', and 'C'. These links are followed by three multiple choice questions, which ask contributors to answer with the label of the clock which comes first, second, and then third chronologically. Additionally, we relied on CrowdFlower's automated testing system to reject workers. We manually filled out the answers to several test questions, which CrowdFlower interspersed with our other HITs. Workers with less than 50% accuracy on the test questions were automatically prevented from completing more tasks.

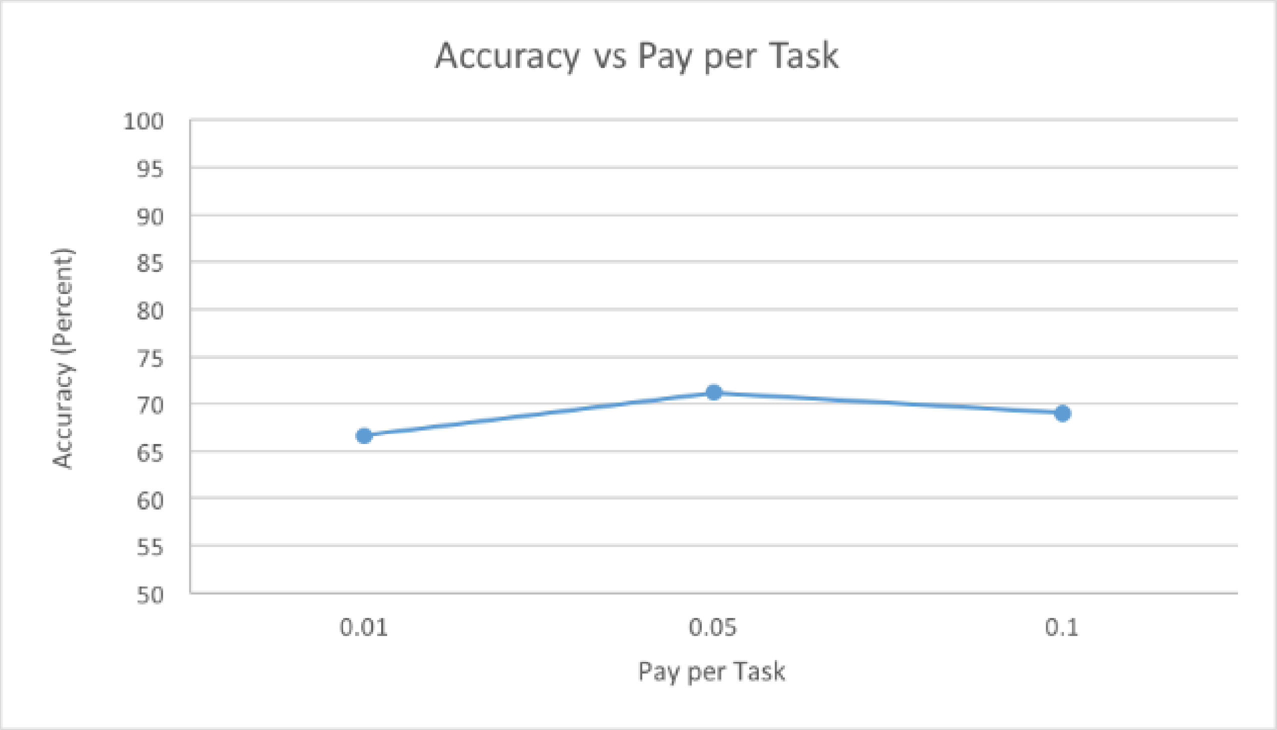
VI. Results

Figure 1



The first metric we measured from the results of our experiment was the number of tasks each worker completed. This was easily obtained from the workers report generated by the Crowdflower platform. We found that with an increase in pay tier, the number of tasks each person completed decreased. Interestingly, our results disagree with the findings of the original paper. However, we attribute this discrepancy to the way our experiment was carried out. For the $0.05 and $0.10 tasks, a much lower percentage of our total available hits were completed because we actually ran out of funds. 100% of the $0.01 tasks were completed, whereas a much lower percentage for the other two were completed. We spent about $50 of our account money for the higher two pay tiers, and we ran out of money pretty quickly. Therefore, the decreased in tasks completed was simply due to the limitations of our experimental design, not as a correlation with the pay tier. If we had enough funds to run the project to completion, we would expect the tasks completed to increase with compensation, mirroring the findings of the paper.

Figure 2



Shown in Figure 2, the results of our experiment are very similar to the findings of the original “Financial Incentives” paper. Our accuracy measurement was obtained by the number of correct orderings over the total number of orderings for each pay tier, checked against the answer key used by the validator script. For the $0.01 pay group, the accuracy was about 66%, which is very close to the 75% accuracy obtained by the original paper. However, when we increased the pay to $0.05 and $0.10, we saw a slight increase in accuracy to 71% and 69%, respectively. We speculate that these two percentages are not statistically significant, but the 5% increase from the $0.01 seems to be. When contributors are completing the hits, the very low pay is an incentive to get through as many as possible, sacrificing accuracy. The reason for this deviance from the original paper might stem from the experimental design of our paper. Our interface was much harder to use, since users had to copy and paste each distinct URL three times for each hit. For such a low pay tier, the contributors for our experiment probably prioritized speed more so than the contributors in the original paper. Overall however, improving the pay level did not lead to an increase in accuracy. This confirms the high-level finding from the first paper.

![HIT Design](https://dl.dropboxusercontent.com/u/17481094/Screen%20Shot%202016-02-25%20at%204.24.55%20PM.png)

https://dl.dropboxusercontent.com/u/17481094/Screen%20Shot%202016-02-25%20at%204.24.55%20PM.png

![Accuracy vs Pay Per Task](https://dl.dropboxusercontent.com/u/17481094/Screen%20Shot%202016-02-25%20at%205.15.51%20PM.png)

![Num. Tasks Completed vs. Pay Per Task](https://dl.dropboxusercontent.com/u/17481094/Screen%20Shot%202016-02-25%20at%205.15.51%20PM.png)