Dylan Mann, Graham Mosley, David Cao, Santi Buenahora
NETS 213
Chris Callison-Burch
2/26/16

Replicating Science Final Report

Section 1: Literature Review

Review 1: *Labeling Images with a Computer Game*

This study used a guessing game to label images. They implemented a Java applet game where two people would be matched in a game. In the game, each player is prompted with the same image, and they must give words that describe that image. However, there is a list of "taboo" words, which is determined from previous labels. Each game gives the players 2.5 minutes to agree on as many images as possible. The researchers then used the words that were agreed on as labels for those images. They gathered up to six labels for each image. Precautions were taken against cheating, such as matching users with differing IP address and matching players with prerecorded "good" games.

Over 13,630 unique people played the game, making 1,271,451 labels for 293,760 images. Once they gathered the labels, they tested the accuracy and relevance of the labels. They also asked crowd workers how many of the labels they would use to describe the image to someone who couldn't see it, and the mean was 5.105 for images with 6 or more labels. Moreover, they asked how many labels didn't make sense with the image, and the mean was 0.105 showing that the labels were very accurate. The researchers concluded that this method is an effective and novel way of labeling images, compared to current methods. Incredibly, the researchers predict that if 5000 players played the game for 24 hours a day, they would label all images indexed by Google in just a couple of weeks.

Review 2: *Financial Incentives and the Performance of Crowds*

This study was about how pay affected the performance of crowdworkers. The researchers ran two experiments, one that involved sorting images, and one that involved finding words in a word search. In each group, they paid workers on three different tiers. The word search group was further divided into workers who were paid by the word, or by the puzzle (assuming a requisite number of words were found).

The results of the experiment were that when participants were paid per element of effort (such as finding one more word or sorting one more image), their performance improved, and they either sorted more images or found more words in the search. This is consistent with standard economic theory that if pay increases, supply of labor increases. The researchers also noted that the pay had no effect on the accuracy of the worker's labor, whether it was sorting or searching for words. One largely surprising result was that when workers were not paid for their work in the wordsearch, they perceived the value of their work to be higher than the workers in the lowest paid segments of each group. The researchers also

noted that their results were consistent with those of typical studies performed in person in laboratories where participants were paid different amounts for performing a task.

Review 3: *Demographics of Mechanical Turk*

A new demographics study was conducted on the users of Amazon Mechanical Turk. This survey was prompted by the change in amazon's policy to allow payment deposits to international bank accounts. This survey included questions about basic demographics (age, location, household information, etc) and questions about how much time was spent and income was made from mechanical turk. Participants were also asked about why they participate in Mechanical Turk. The survey, conducted in February 2010, contained 1,000 participants who were paid $0.10 each.

The survey found that almost half of the participants were from the United States and another 34% were from India. Since the United States and India were the two most popular countries, many comparisons of different demographic information were made between the two countries. The results of Mechanical Turk involvement questions were also analyzed. The study found that most people work a day or less per week and have made less than $20 total. A significant amount of Indian workers stated that Mechanical Turk was their main source of income, while the vast majority of the Americans surveyed used Mechanical Turk purely as a secondary income.

Section 2: Description and Methodology of *Financial Incentives and the Performance of Crowds*

This study involved two different experiments. Both began by paying participants $.10 to complete a demographics survey posted on Mechanical Turk. The results of this survey were recorded and later used for a variety of reasons. First, it was important to verify that the sample that participated in the survey was representative of Mechanical Turk's demographic. Second, it allowed the results from the later experiments to be analyzed by demographic in addition to on aggregate.

After the survey had been completed the participants did a Human Intelligence Task(HIT). In the first experiment, the HIT evaluated Turkers' abilities to sort images by the number of cars that appeared in each the images. In each HIT, the workers were presented with (the same) initial 3 images, and asked to sort them based on the number of cars in each image. Their performace was compared with the correct result, and the workers were given feedback based on whether they correctly sorted the images. The workers could sort between 2 and 4 images, and were paid for each HIT based on how many they chose to sort. The participants were randomly assigned to be paid either $.01$,.05, or $.10 per image sorted, and then the performance of each of the three different price groups was evaluated for correctness.

The second experiment evaluated the quality and quantity of words participants found in a word search. Each participant was given a list of 15 words that could potentially be in the wordsearch. Of those 15, more than 8 actually belonged. These participants were also broken into two groups, whether they were paid per word or for a finding a certain quota of words. Each of these three groups had 3 pay scales, and were compared to a baseline where participants were not paid at all. The quota groups were paid either $.01$,.05, or $.10 per puzzle completed, and the per word groups were paid either .01, .02, or .03$ per word.

The results of the both studies were interesting. In the image sorting study, workers completed more image sorting activities when pay was higher. However, they also found that there was no significant difference in the accuracy of the sorting done by workers. This is counter to what they expected, as they expected both more work and better work, but the workers only increased quantity, not quality.

In the puzzle solving study the researchers also found that effort-contingent pay motivated participants to do more work. Participants who were paid either on a quota or a piece-rate basis completed more puzzles and found more words per puzzle than participants who were not paid. They also noted a positive correlation between amount of pay and the number or words found and puzzles solved. They also noted that participants who were paid per word earned 4 times as much as those that were paid per puzzle (on a quota basis). In spite of this difference, participants paid per puzzle found far more words than those paid per word. This is a counterintuitive result, that the researchers attributed to the marginal benefit to finding the last few, difficult to find words. For per word pay, participants may be incentivized to move on to the next puzzle before finding the hardest words, whereas quota participants would need to stay on a puzzle and finish the quota if they wanted to be paid at all.

Section 3: Our Experiment Methodology

We decided to recreate the second experiment from *Financial Incentives and the Performance of Crowds*, where workers were given a demographics survey, then approved for a HIT where they were paid per word found in a word search. First we posted 300 demographics survey HITs. This HIT asked questions about the location, age, gender, education, ethnicity, and income level. This information was aggregated and upon completion of this HIT, a boto script, named " granted one of 3 qualifications to the crowdworker. These qualifications pertained to one of three word search HIT groups with 100 HITs each. All three batches had the same rate of base pay, ($.01) however one paid an additional \$.01 per word, one paid .02$ per word, and one paid \$.03 per word. These bonuses were granted by a boto script that was run periodically (included in our additional files, called grant_bonus.py). An additional qualification was granted to each of the workers upon completion of the survey that prevented them from retaking the survey.

Each of the word search HITs contained one of 100 unique wordsearches that was generated locally using a Python script, named MakeWordSearch.py. The word searches we generated were in the same format as the ones in the original study, a 15 x 15 puzzle with 15 words that could possibly be in the puzzle, and 8-12 that were actually in the puzzle. Words were between 5 and 8 letters long, and were

selected from the Windows system dictionary.

In the HIT, words could be selected by clicking the beginning of a word and dragging and releasing until the end of the word was found. If the selected word was not correct, then it did not stay highlighted, but if it was correct, the word on the side of the puzzle was crossed out, and the highlighted word stayed until the puzzle was submitted. The backend of the HIT was written in JavaScript, and the words found were listed as a hidden input field in the HTML of the HIT. This allowed the information to be submitted along with the rest of the HIT, and bonuses could be granted accordingly. At the bottom of the wordsearch was an input field where respondents were asked how much they thought their effort was worth (regardless of pay).

We decided to use Amazon Mechanical Turk for our experiment, for a number of reasons. First, it minimized the deviation from the original experiment, as the original also used MTurk for its experiments. Secondly, Mechanical Turk provides customization and automation tools that far outperform anything available on CrowdFlower. Boto scripts allowed the assignment of bonuses and qualifications that would not have been nearly as easy using CrowdFlower's tools. The amount of workers on MTurk also far outnumbers the number of workers on CrowdFlower, allowing the batch to complete faster.

Each of the initial survey HITs paid $.10, and each wordsearch hit paid a .01$ base reward and a bonus per word (as stated above). Workers were given 30 minutes to complete the survey HITs and 3 hours to complete each wordsearch HIT. We decided to reject anyone who completed the survey more than once (there were explicit instructions to only complete it once), and also anyone who submitted zero words in a word search (it was out of principle, most spent less than 20 seconds searching). The HTML and JavaScript for the wordsearch HITs can be found in the included file wordsearchhit.html. The survey HIT can be observed in the included file surveyhit.html. By the conclusion of our HITs, 293 people had taken the survey, and 127 had completed at least one puzzle.

Section 4: Deviations from the Original Study

We deviated from the original experimental design in several ways over the course of our study. We decided to only implement the word search task experiment, not the image sorting one. The main motivation for this change was that we did not have an effective way of sorting the images originally. This meant that analyzing the results of the workers' accuracy would have been very difficult. Additionally, we did not have a good source of images to use, whereas we could fairly easily create a word search maker using Window's system dictionary and a Python script. Furthermore, we opted to only implement the pay-per-word style of pay for solving puzzles, rather than the quota payment method as well. This was due to constraints on money, because implementing both methods would have required either half the sample size (and worse data), or double the expenses. Seeing as we are poor college students, but wanted good results, we decided to collect good data on just one of the experiments.

We also made a few changes to the HIT design. The original study used one HIT per participant that included an opening survey, as many wordsearches as the respondent wanted to solve, and an exit survey. Instead, we decided to assign one of 3 qualifications, that permitted the participant to complete a corresponding word search HIT (identical aside from 3 different rewards per word). We also prevented the participant from completing a survey again by assigning a "Survey Taken" qualification. The reasoning behind this change was that we could not figure out a good way of simultaneously notifying the respondent of their pay, while also preventing them from taking differently paying HITs. One alternative was to only post certain paying hits at any given time, but with the two week deadline, and the week that it took to design the hit and create the wordsearches, there was too little time to run HITs over different timeframes. The alternative method would have also potentially caused certain people to complete two word search HITs at different prices per word.

We also decided not to implement the no pay puzzle. This was because our puzzles were in separate hits from our survey, contrary to the original study, so there would have been no visibility for our free to solve wordsearches, or if we paid a base rate of, maybe a cent, then workers would just submit empty puzzles to cheat the system.

We also changed the number of HITs from the original study. The study had 320 participants, who completed a total of 2736 puzzles, finding 23,440 words. We determined that these sample sizes would be prohibitively expensive for us to duplicate. Therefore we put out 300 survey HITs (of which 7 were taken by duplicates, for 293 total participants) and 300 total word search HITs, 100 of each payment model. Of the 293 survey takers, 127 went on to complete at least one word search (and 3 submitted empty word searches). This is a poor clickthrough rate, but considering that our experimental design allocated one word search per survey participant, it was not a problem that so few of the survey respondents completed puzzles.

One final deviation from the original study was that we did not require participants to take an exit survey after completing wordsearches. This was also due to the logistics of using Mechanical Turk. Our design was already complicated, utilizing qualifications and different HITs for each level of pay, and so we decided that it would confound our data if we used an exit survey as well (either separate from the wordsearch hit or after each completed puzzle). Instead, we just asked each participant one question after they were done with their wordsearch, "Disregarding how much this HIT is paying, how much do you think each word was worth (from $.01 to .10$)?"

Section 5: Results and Discussion

The demographics breakdown for our experiment was significantly different from the demographics in the original experiment. The largest change was that rather than 74.4% of respondents being women (as they were in the original study), only 49.5% of the survey respondents to our survey were women. However, of the 127 people who solved puzzles, 57% were women, suggesting some gender bias in our results as well. Some of the survey results may have been completed incorrectly though, such as the

American 8 year old girl who had already obtained a masters degree (in spite of our form being constrained to a minimum input of 12), and the multiple 12 year old children who had already received high level degrees (one each of PhD, Masters, and Bachelors degrees). Our best guess is that their degrees didn't get them good jobs ($7000–30,000 per year), and they needed to complete wordsearches to make ends meet.

Our participants found a total of 2218 words. Interestingly, they found an average of 7.39 words per puzzle, which is more than a full word less than the original study's results of 8.57 words per puzzle. It is hard to determine what the source of this discrepancy was, however, the original study was composed of 74% women, and the researchers speculated that their sample was skewed because their workers enjoyed solving the puzzles. Perhaps our data was more indicative of normal workers completing tasks, rather than enjoying word search puzzles. The paper had no information about the distribution of words per puzzle, so we guessed it was 8-12, and maybe there were fewer words to find overall in our puzzles.

One of the most astounding results we found was the difference between men and women, as shown in the graph below.

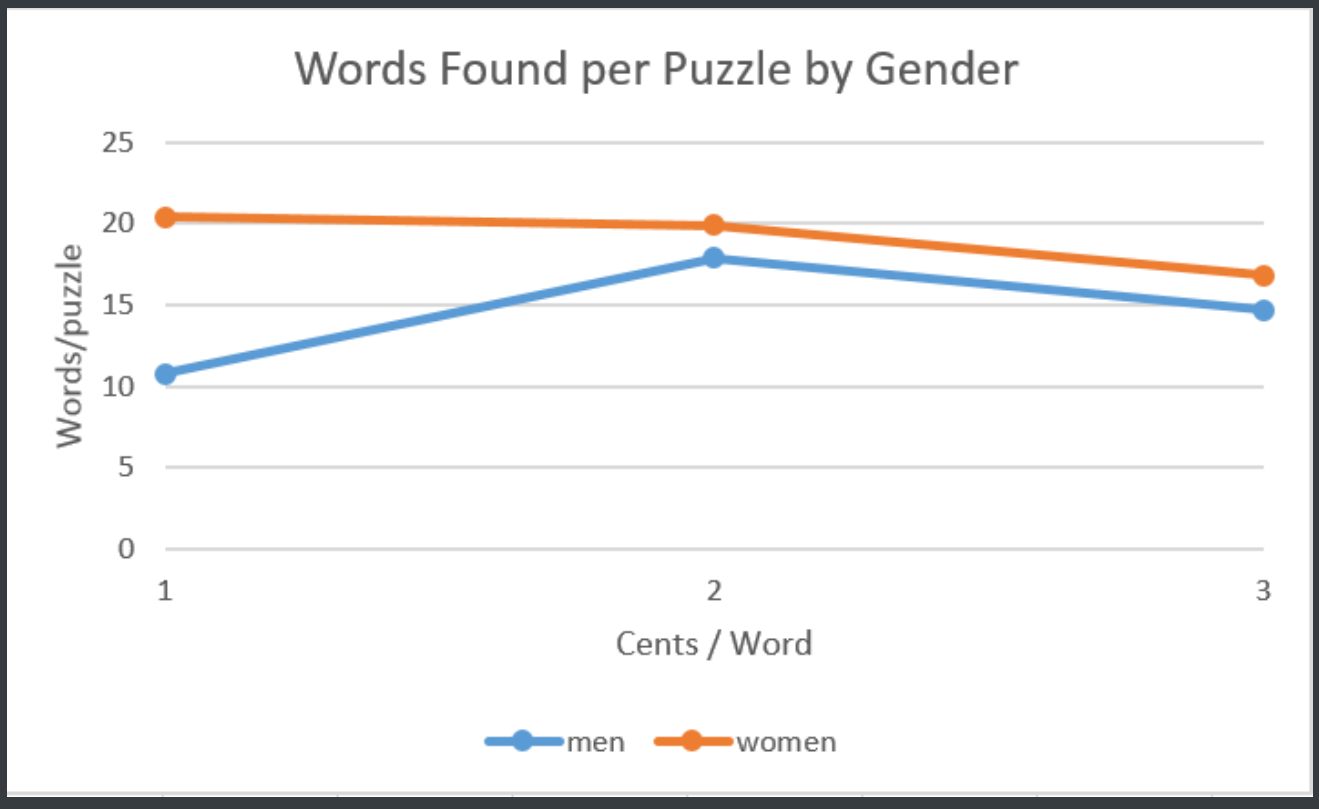

Figure 1: words found per puzzle by gender

When pay was only 1 cent per word, woman found twice as many words per person as men. This gap narrows as pay increases, and our sample sizes were small, but that is still an extremely significant difference. These results were similar to those of the original study, and the researchers speculated that this was due to more men on MTurk being motivated financially, whereas women would just stick with a task if they enjoyed it.

Another interesting trait of this data is in regards to the relationship between reward and resulting rate of words per person. We found that workers paid 2 cents per word found more words per person than workers who were paid 1 cent per word, but this trend did not hold when pay was increased to 3 cents per word. One explaination for this is that workers who were paid more per word were thinking more about the monetary reward, and less about enjoying the puzzles. One piece of evidence for this is that from 1 cent to 2 cents, workers spent twice as long on puzzles, but they spent about the same amount of time on 2 cent and 3 cent per word puzzles. This data was different from the original study, that found that increasing pay always increased work done.

Finally, if we consider perceived value per word (the chart shown below), we see that crowd workers feel like their work is worth more money when the reward per word is higher, suggesting that the original study's observation of an "anchor" based on task pay is correct. However, one change was that perceived value per word did not change when pay changed from 1 cent to 2 cents. Perhaps this was due to our slightly different question wording, which gave a possible range from 1 - 10 cents for the perceived value of the task.

Perceived Value vs Reward

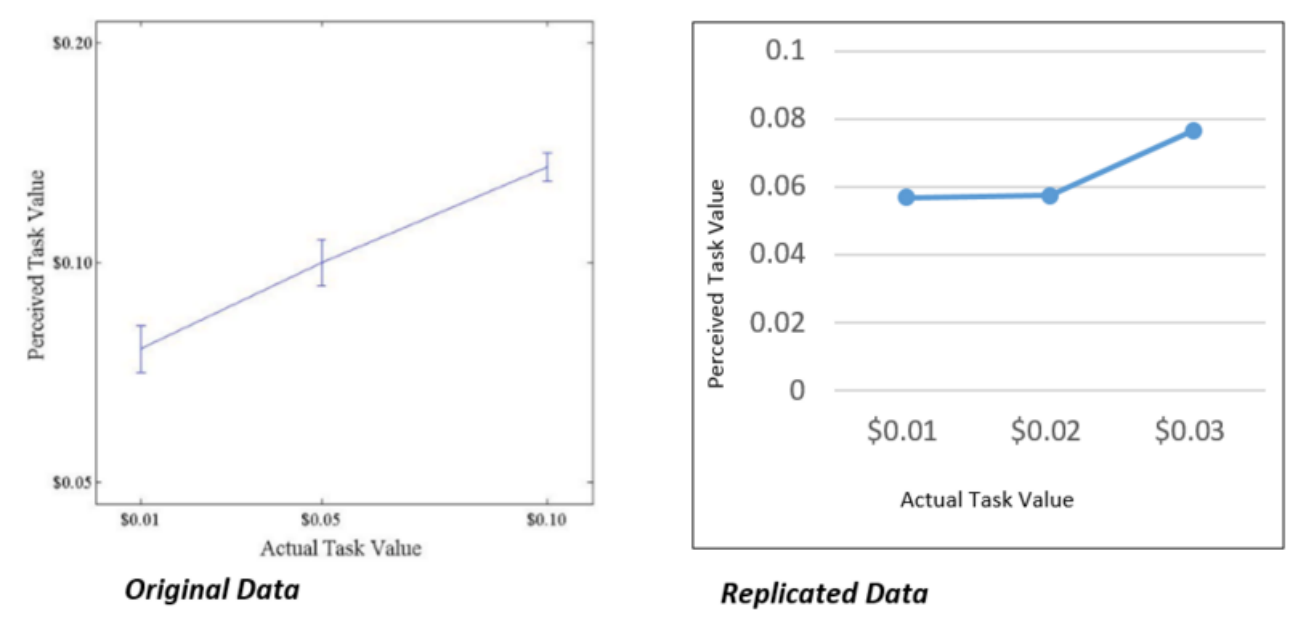

Figure 2: Perceived value of task vs. reward.

Their argument was that this participant likely continued working out of enjoyment. Our data speaks of similar trends, where one worker spent the full 3 hours on a $.01 puzzle, and several spent 3 hours working on .03 cent puzzles.

Our calculated 'per hour' rates suggest that a Turker working continuously for an hour could expect to earn between $.30 an hour for 1 cent/word puzzles and 42 per hour for 3 cent/word puzzles. This is interesting because we used exactly the same 100 puzzles in each of the 3 HIT batches. and yet workers only made slightly more per hour when pay per word increased, due to finding more words and spending

more time on each puzzle. This suggests that workers put in time and effort that scaled with the pay increases, contrary to the original results. One particularly ambitious worker spent 5 hours working on 3 puzzles, making a total of $.25 (but finding all of the possible words!).

In conclusion, our experiment shared many similarities with the original. Both our study and the original observed a positive correlation between pay and amount of work. Both studies also observed a lot of noise in the data due to people enjoying puzzles instead of caring about compensation. We found that quality of the work (or at least effort) did increase with pay, contrary to the original findings. Worth noting is that our small sample size could have affected our findings, but nevertheless, it is clear that Mason and Watts' findings were reflected in many parts of our experiment.