Quality Control - part 2

Crowdsourcing and Human Computation

Instructor: Chris Callison-Burch

Website: crowdsourcing-class.org
Different Mechanisms for Quality Control

• Aggregation and redundancy
• Embedded gold standard data
• **Economic incentives**
• Reputation systems
• Statistical models
Does pay impact quality?

- Economic theory holds that workers are rational actors.
- Will choose to improve their performance in response to a scheme that rewards improvements with financial gain.
- Example: executive compensation tied to stock price.
Different pay schemes

• Lazear studied of workers who installed windshields on a production line

• Switched from pay per hour to pay per unit during a year and a half

• Individual productivity for workers who started in the hourly rate and switched to the per-unit scheme increased by 20%

• Conclusion: performance-based pay schemes can elicit improved performance
Is that the whole story?

• Sometimes financial incentives can undermine “intrinsic motivation”. This can lead to poorer outcomes.

• For complex tasks, performance pay can encourage workers to focus only on the aspects of their jobs that are actively measured.

• Can also lead to employees avoid taking risks, thereby hampering innovation.
Financial Incentives and the “Performance of Crowds”

• Experiment with economic incentives on Amazon Mechanical Turk
• An exciting tool for behavioral research, since you can recruit thousands of participants from a real labor market
Impact of compensation

- Does compensation change the quantity of work performed (output)?
- Does it change the quality of the work (accuracy)?
Participants were informed that the feedback would only be correctly sorted, and if not, what the correct order was. They were given a chance to practice sorting images. When they felt the images they chose to finish sorting image sets, or if they sorted all 99 image sets, they were asked to complete a brief introductory survey and accept the bonus (if any) for the tasks so far completed. Once their work to receive payment.

Participants could sort any number from 0 to 99 sets of images, and for others there was no such indication. If examined further, there was no indication that the difficulty or payments, either the manne-

One might therefore suspect that participants in our experiments associated with a unique ID (which is, in turn, associated with a wish to differentiate between the quantity of work performed.

Camera at 2 second intervals into chronological order. To sort an image set for which output can vary widely and accuracy can (output) and the quality of the work (accuracy); thus we require a measure of quality where the number of image sets they chose to sort was our feedback about the number of tasks completed and could submit their sorting and proceed to the drag it into the correct position. When they felt the images, displayed in the same (incorrect) order. To sort an image.

Participants were required to have an Amazon's Mechanical Turk account on Amazon's Mechanical Turk. These accounts are associated with a unique ID (which is, in turn, associated with a wish to differentiate between the quantity of work performed.

One might therefore suspect that participants in our experiments associated with a unique ID (which is, in turn, associated with a wish to differentiate between the quantity of work performed.

The second issue is that payments in crowd experiments, and trivial compared with traditional labor markets. One might therefore suspect that participants in our experiments associated with a unique ID (which is, in turn, associated with a wish to differentiate between the quantity of work performed.

Over all conditions, the experiment involved 611 participants, 47% were female and 53.41% reported an income between $7000 and $70k, 22.9% reported an income less than $7000, 22.9% reported an income between $70 and $30k, and 22.9% reported an income greater than $30k. The median age was 25 years old. The majority of participants were from the United Kingdom (62.2%), followed by the United States (18.2%).

Participants could sort any number from 0 to 99 sets of images, and for others there was no such indication. If examined further, there was no indication that the difficulty or payments, either the manne-

3.1.1 Design

We randomly assigned participants to one of four pay levels: no pay, low pay ($0.05 per image set), medium ($0.10 per image set), and high pay ($0.20 per image set). These payments are much smaller (cents per task) than would typically be the case in lab experiments, and trivial compared with traditional labor markets.

The first issue is that workers will not respond in a sensible way to incentives because they are motivated primarily by non-expert

suggesting that there is some external validity to the effects.
Payment scheme

• Everyone: $0.10 for doing training examples and filling out a survey
• Payment levels: nothing, 1¢, 5¢, 10¢ per set
• Num images per set (independent of payment): 2, 3, 4
• Each person sorted up to 99 sets of images, could end participation at any point and get paid for what they did
• 611 subjects sorted a total of 36,425 image sets
Number of tasks done

The number of tasks done in different ways: first, using a simple one confirmed quantitatively what is visually apparent in Figure 3 in order and the sorted order. For each accuracy, the normalized sum of squared differences between the correct order; and second, using Spearman's rank correlation (the proportion of image sets that were sorted into the correct output on compensation therefore suggests that the range of wage rate would have overwhelmed the effect of enjoyment of the task) would have overwhelmed the effect of changes in extrinsic motivation (payment), which can vary by at.

As Figure 3 indicates, however, increasing compensation did not decrease with increasing difficulty. We also observe, however, that there is no interaction between difficulty and compensation, decreases with difficulty levels. Although difficult levels participants chose to complete more tasks on average when the pay was higher (Figure 3 reveals two main findings: first, that across all difficulty levels participants chose to complete more tasks on average when the pay was higher (Table 1.

Results

Average parameter estimates for the effect of pay in the hierarchical linear model across users.

```
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0.059</td>
<td>0.029</td>
</tr>
<tr>
<td>β</td>
<td>0.029</td>
<td>0.055</td>
</tr>
<tr>
<td>σ²</td>
<td>0.220</td>
<td>0.055</td>
</tr>
</tbody>
</table>
```

Figure 2.

```
<table>
<thead>
<tr>
<th>Number of Tasks Completed</th>
<th>Pay per Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$0.00</td>
</tr>
<tr>
<td>2 images</td>
<td>30</td>
</tr>
<tr>
<td>3 images</td>
<td>20</td>
</tr>
<tr>
<td>4 images</td>
<td>10</td>
</tr>
</tbody>
</table>
```

Table 1:

Subject pool was therefore reasonably diverse, consistent with demographics.html-2 reveals two main findings: first, that across all difficulty levels participants chose to complete more tasks on average when the number of tasks each participant completed.

```
F(3,607) = 15.73, p < 0.001
```

Analyzed at the participant level, the measures of accuracy can be analyzed at the task level; thus, the hierarchical linear model because it accounts for the variability in the number of tasks each participant completed.

In this analysis, the variable inherent difficulty of sorting each image set, and task slope for the wage received, and also found that more of the participants paid $0.10 sorted the maximum possible than those paid $0.01 or nothing at all, and also recorded 58% female respondents.

Nevertheless, the finding is reassuring since, as noted above, one more a person is paid to do X, the more of X they will do consistent with standard economic theory, which predicts that the proportion of image sets sorted in correctly, the model is that the image set...error bars are standard error.

```
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0.059</td>
<td>0.029</td>
</tr>
<tr>
<td>β</td>
<td>0.029</td>
<td>0.055</td>
</tr>
<tr>
<td>σ²</td>
<td>0.220</td>
<td>0.055</td>
</tr>
</tbody>
</table>
```
as if you were reading it naturally.
Regardless of the accuracy measure or analytical method used, we found that the wage rate had no significant effect on the participants' accuracy in sorting the image sets. First, as indicated in Table 1, the parameter estimates in the hierarchical model for the four levels of pay were not reliably different from each other; and second, one-way ANOVAs of wage rate on proportion correct and rank correlation were not statistically reliable (proportion correct: $F(3,607) = 0.66$, ns; rank correlation: $F(3,607) = 0.82$, ns).

3.3 Discussion

One possible explanation for the absence of an effect of wages on accuracy is that subjects simply assumed they would be paid regardless of performance. This explanation is somewhat unlikely, as AMT's policy is that requestors are only obligated to pay for accurate or useful work, and workers are informed of the policy. Nevertheless, to check the possibility we ran an additional experiment with a single payment level ($0.01) that provided different information to participants regarding the importance of accuracy. In this additional experiment, some participants were given the same instructions as before while others were told that one out of every four image sets was a test image set used to gauge their accuracy. Within this latter condition, we also created four variants: (i) participants only informed that accuracy would be measured; (ii) participants also shown feedback on their accuracy after every fourth image set; (iii) participants also told explicitly that their pay would be contingent on their performance; and (iv) participants shown feedback and also told that pay was contingent. We found that quantity and quality results were indistinguishable in all these conditions, suggesting that participants in all conditions were in fact treating their pay as performance dependent.

Although the differential effect of pay on quantity and quality is at first puzzling, we note that previous studies have also found positive effects of financial incentives on quantity of work performed but no effect on quality [24]. We hypothesize, moreover, that the difference derives from an "anchoring" effect, similar to effects that have been observed in other domains of judgment and decision-making [19-21]. As Figure 4 shows, when surveyed after the completion of their tasks, workers in all conditions generally felt that the appropriate compensation for the work they had just performed was greater than what they had received, but the values they expressed depended significantly ($\chi^2 = 243.61, p < 0.0001$) on their actual compensation: on average, workers paid $0.01 per task felt they should have received $0.05; workers who were paid $0.05 felt they should have received $0.08; and workers who were paid $0.10 felt they should have received $0.13. On the one hand, therefore, paying people more to perform a task makes that task more attractive relative to their available outside options, such as other HITs on AMT; thus subjects in the higher pay conditions stayed longer and completed more tasks than those in low pay condition. On the other hand, because of the anchoring effect, all workers felt like they were being paid less than they deserved; thus were no more motivated to perform better no matter how much they were actually paid.

4. STUDY 2: WORD PUZZLES

4.1 Methods

In spite of this explanation, one might suspect that the absence of an effect on accuracy may be an artifact of the task itself—because, for example, it allowed only a small number of potential solutions (in the "easy" condition, for example, only two solutions were possible); or because subjects could not easily improve the quality of their answers with greater effort. To address this possibility, we performed another experiment, using a similar experimental design, but changing the task to finding words hidden in a random array of letters (see Figure 5).
Word Jumble Puzzles

- Find as many of the words in a set as you can:
- ACHIEVE, ATTAIN, BUILDING, CHAIR, COMPLETE, GREEN, LAMP, MASTER, MUSIC, PLANT, STAPLE, STEREO, STRIVE, SUCCEED, TURTLE
- Not all of the words listed are in the puzzle!
Experimental setup

- Different pay rates (just as before)
- Subjects were told that they would be paid either on a per-grid basis or a per-word basis, or not told anything
- quantity = number of puzzles completed
- quality = fraction of words found per puzzle
- Participants could do up to 24 puzzles
- 320 subjects solved 2736 puzzles, finding 23,440 words
Participants would encounter one of two practice puzzles, each comprised of 15 words, some of which were given more detailed instructions about how to do the task. If participants accepted the task, they were given feedback on their performance and could submit the puzzle when they chose to finish, or if they completed all 24 puzzles, they were paid for every word found. Within each compensation scheme, we once again created three pay levels. For each puzzle, there was a list of 15 words, some of which were hidden, and the payment would be by word, after which the word would appear in a panel on the right. To select a word, participants would click on the start and end of a word, and the word would appear in a panel on the right.

In Experiment 2, we once again created three pay levels—fun, pay per puzzle, and pay per word. Within each compensation scheme, we varied compensation levels, and the level of compensation was not varied for the word. The recruitment of participants was the same as in Study 1, with a concern that as we show next, there was a very strong relationship between the number of puzzles completed and the compensation level. There was also a very strong relationship between the number of puzzles completed and the compensation level. The recruitment of participants was the same as in Study 1, with a concern that as we show next, there was a very strong relationship between the number of puzzles completed and the compensation level. The recruitment of participants was the same as in Study 1, with a concern that as we show next, there was a very strong relationship between the number of puzzles completed and the compensation level.
other words, those in the pay
completing the puzzle becomes, in effect, a more salient goal). In
puzzle; and second, through implicit goal setting (i.e. where
quota may elicit work in two ways: fi
payment). Following this work, we note that the presence of a
has found that quota systems (analogous to our per
Although counterintuitive, the higher work
found the least (81.4%).

paid per puzzle found the next highest (84.7%), and those paid per
line) shows, however, the actual ranking was precisely the
more words than those not being paid at all. As Figure 7 (dashed
puzzle than those being
that participants being paid per
scheme itself had
words found per puzzle
compensation had no significant effect on the quality of
it is not surprising that we again found that the le
important than in task 1, it remains relevant to output.
them, suggesting that although extrinsic motivation play a less
paying participants to work generated more work than not paying
the task intrinsically enjoyable, thus diminishing the impact of
it a little. Presumably, therefore, at least some participants found
in the hierarc
$0.$
$0.0$
$0.10$
$0.05$
$0.01$

"pay
shows the per
per puzzle; and B) paid per word.
measured here as the fraction of total possible
word equivalent pay for the three payment
Model estimate
—
95% Confidence Interval

-0.036 $0.03$
-0.022 $0.01$
-0.01 $0.00$
0.0 $0.0
0.01 $0.00$
0.027 $0.00$
0.0357 $0.00$
0.0464 $0.00$

—
Data to a hierarchical linear model across users when A) paid
-puzzle condition may have
-0.066,
-0.027,
-0.01,
-0.03,
-0.025,
-0.03,
0.0,
0.0
0.0
0.0
0.0
0.0
0.0

-0.03 $0.03$
-0.025 $0.025$
-0.02 $0.02$
-0.015 $0.015$
-0.01 $0.01$
-0.005 $0.005$
0.005 $0.005$
0.01 $0.01$
0.015 $0.015$
0.02 $0.02$
0.025 $0.025$
0.03 $0.03$
0.03 $0.03$
0.03 $0.03$
0.03 $0.03$
Our results are silent differences between observing any such effects. We randomly assigned individuals to different pay conditions, we cannot employ employers to attract and retain better workers. Because we allocate much of their impact through sorting effects, for instance, it may be the case that financial incentives exert financial incentives systems that have been studied elsewhere. Because our experimental design skirts some important aspects of the subject pool and pay rate issues raised earlier, the context of crowd sourcing is more speculative, in part because of the diminished quantity of individual output that the low pay would yield, but not better. Assuming, of course, a large enough crowd exists to make up for the diminished quantity of individual output the low pay would yield, but not better. The motivation (i.e., enjoyable work) appears to be robust, suggesting this important issue of anchoring is not sensitive to the specific nature of the task (e.g., the actual pay and perceived value). Although a number of our results do seem sensitive to the specific nature of the task (e.g., the dependency of quantity on pay rate), the anchoring effect seems to be general. In addition to these real world analogs, we have already noted that dependency of quantity on pay rate, this anchoring effect seems to be quite general. In particular, it has been shown that financial rewards, such as those obtained in physical laboratories where the sums of money at stake were considerably larger than in our case (albeit still small compared with pay in traditional labor markets). In particular, it has been shown that financial compensation tends to increase systematic relation to compensation, and in part to understanding these results, we note that absolute pay rates are less important to performance than the kind we have described here; they are needed to make firm judgments about performance. Rather, it appears to derive because Europeans are paid considerably less than their counterparts, and in part from historical differences in cultural norms, which have one provocative analog is the observation that chief executives in the US receive the lowest contingent pay, participants had no anchor and perceived the quality of the work will be when they are not being paid anything, they have nothing to judge the value of their work. Analogous to what we see in our results reported in a post-observational analysis of the kind we have described here would be needed to make firm generalizations, therefore, these results do at least suggest that the effect of setting an “anchor,” relative to which individuals are judging the value of their work, can be accomplished as effectively for little to no cost. Second, as good or better than using financial rewards, it may be that it is possible to use non-financial rewards, in other words, may get the work done faster, but not better. When not given task contingent pay, participants had performance similar to the highest (left) or a word (right) performance may be systematic relation to compensation, and in part to understanding these results, we note that absolute pay rates are less important to performance than the kind we have described here; they are needed to make firm judgments about performance. Rather, it appears to derive because Europeans are paid considerably less than their counterparts, and in part from historical differences in cultural norms, which have one provocative analog is the observation that chief executives in the US le workers in the US document pay gap between US Federal circuit judges found that absolute pay rates are less important to performance than the kind we have described here; they are needed to make firm judgments about performance. Rather, it appears to derive because Europeans are paid considerably less than their counterparts, and in part from historical differences in cultural norms, which have one provocative analog is the observation that chief executives in the US.
Findings

• Paying subjects elicited higher output than gamification, and increasing pay rate yielded even higher output.

• However, paying subjects did not affect their accuracy.

• Anchoring effects are significant – the reward you set impacts perceived value.
Implications for your tasks?

• When you can use non-financial rewards, like intrinsic motivation, do so, since the quality of work will be the same.

• When you can’t use intrinsic motivation, it might be in your best interest to pay as little as possible. Your work will be done slower, but quality will be similar.

• Is this fair to workers?
What do you think?

- Is studying workers on Mechanical Turk a valid way of studying other labor markets?
- What possible confounds are there?
- What could we do to control for them?