Crowdsourcing and Human Computation

Instructor: Chris Callison-Burch

Website: crowdsourcing-class.org
What will we cover in this class (and should you take it)?
Syllabus

- Taxonomy of crowdsourcing and human computation
- The Mechanical Turk crowdsourcing platform
- Programming concepts for human computation
- The economics of crowdsourcing
- Crowdsourcing and machine learning
- Applications to human computer interaction
- Crowdsourcing and social science
Who should take this class

- Anyone who wants to be on the cutting edge of this new field
- Entrepreneurial students who want to start their own companies
- Students from the business school who want to experiment with markets
- Students from the social sciences who want to conduct large-scale studies with people
What will you get out of this class?

• Understanding of an emerging field of CS
• Basic python and machine learning skills
• Ideas that you could transform into a startup company or academic research
• A new way of thinking about collective decision making companies and countries
Inter-related concepts

Collective Intelligence

“Groups of individuals doing things collectively that seem intelligent”

Human Computation

“A paradigm for utilizing human processing power to solve problems that computers cannot yet solve.”

Crowdsourcing

“Outsourcing a job traditionally performed by an employee to an undefined, generally large group of people via open call.”

The Sharing Economy

“An economic system in which assets or services are shared between private individuals, either for free or for a fee, typically by means of the Internet.”

Data Mining

“Applying algorithms to extract patterns from data.”
Crowdsourcing Companies

“Outsourcing a job traditionally performed by an employee to an undefined, generally large group of people via open call.”
Crowdsourcing Companies

“Outsourcing a job traditionally performed by an employee to an undefined, generally large, group of people via open call.”
Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it’s convenient.

37,649 HITs available. [View them now](#).

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now](#).

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Register Now](#)

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you’re satisfied with the results

[Find HITs Now](#)

or [learn more about being a Worker](#)

or [learn more about being a Requester](#)
Rewards over past 5 years
## Top Requesters

<table>
<thead>
<tr>
<th>Requester ID</th>
<th>Requester Name</th>
<th>#HIT groups</th>
<th>Total HITs</th>
<th>Rewards</th>
<th>Type of tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3MI6MIUNWCR7F</td>
<td>CastingWords</td>
<td>48,934</td>
<td>73,621</td>
<td>$59,099</td>
<td>Transcription</td>
</tr>
<tr>
<td>A2IR7ETVOIULZU</td>
<td>Dolores Labs</td>
<td>1,676</td>
<td>320,543</td>
<td>$26,919</td>
<td>Mediator for other requesters</td>
</tr>
<tr>
<td>A2XL3J4NH6JI12</td>
<td>ContentGalore</td>
<td>1,150</td>
<td>23,728</td>
<td>$19,375</td>
<td>Content generation</td>
</tr>
<tr>
<td>A1197OGL0WOOQ3G</td>
<td>Smartsheet.com Clients</td>
<td>1,407</td>
<td>181,620</td>
<td>$17,086</td>
<td>Mediator for other requesters</td>
</tr>
<tr>
<td>AGW2H4I480ZX1</td>
<td>Paul Pullen</td>
<td>6,842</td>
<td>161,535</td>
<td>$11,186</td>
<td>Content rewriting</td>
</tr>
<tr>
<td>A1CTI3ZAWTR5AZ</td>
<td>Classify This</td>
<td>228</td>
<td>484,369</td>
<td>$9,685</td>
<td>Object classification</td>
</tr>
<tr>
<td>A1AQ7EJ5P7ME65</td>
<td>Dave</td>
<td>2,249</td>
<td>7,059</td>
<td>$6,448</td>
<td>Transcription</td>
</tr>
<tr>
<td>AD7C0BZNKYGYV</td>
<td>QuestionSwami</td>
<td>798</td>
<td>10,980</td>
<td>$2,867</td>
<td>Content generation and evaluation</td>
</tr>
<tr>
<td>AD14NALRDOSN9</td>
<td>retaildata</td>
<td>113</td>
<td>158,206</td>
<td>$2,118</td>
<td>Object classification</td>
</tr>
<tr>
<td>A2RFHBFTZHX7UN</td>
<td>ContentSpooling.net</td>
<td>555</td>
<td>622</td>
<td>$987</td>
<td>Content generation and evaluation</td>
</tr>
<tr>
<td>A1DEBE1WPE6JFO</td>
<td>Joel Harvey</td>
<td>707</td>
<td>707</td>
<td>$899</td>
<td>Transcription</td>
</tr>
<tr>
<td>A29XDCCTJMAE5RU</td>
<td>Raphael Mudge</td>
<td>748</td>
<td>2,358</td>
<td>$548</td>
<td>Website feedback</td>
</tr>
</tbody>
</table>
A few requesters offer most of the rewards
HITs by price
<table>
<thead>
<tr>
<th>HIT Type</th>
<th>Requester</th>
<th>HIT Expiration Date</th>
<th>Time Allotted</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription of approximately 15 minutes of audio</td>
<td>Amelia Jones</td>
<td>Sep 5, 2013 (2 days 23 hours)</td>
<td>8 hours</td>
<td></td>
</tr>
<tr>
<td>By Invitation Only: Answer a few questions in a brief survey</td>
<td>Qualtrics Survey</td>
<td>Sep 5, 2013 (2 days 21 hours)</td>
<td>30 minutes</td>
<td></td>
</tr>
<tr>
<td>Take a geo-tagged photo of a landmark in North Dakota</td>
<td>Crowdsourcing at Thomson Reuters</td>
<td>Sep 5, 2013 (2 days 16 hours)</td>
<td>3 days</td>
<td></td>
</tr>
</tbody>
</table>
### Manage Batches

Click on the name of the batch to see more details.

**Batches in progress (1)**

<table>
<thead>
<tr>
<th>Batch Name</th>
<th>Created</th>
<th>Assignments Completed</th>
<th>Time Elapsed</th>
<th>Estimated Completion Time</th>
<th>Average Time per Assignment</th>
<th>Effective Hourly Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression HIT - grammar/meaning 10</td>
<td>July 04, 2013</td>
<td>2,468 / 2,468</td>
<td>1 day</td>
<td>Complete</td>
<td>3 minutes 40 seconds</td>
<td>$4.091</td>
</tr>
<tr>
<td>Batch Progress</td>
<td>100% submitted</td>
<td>100% published</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Batches ready for review (143)**

<table>
<thead>
<tr>
<th>Batch Name</th>
<th>Created</th>
<th>Assignments Completed</th>
<th>Time Elapsed</th>
<th>Estimated Completion Time</th>
<th>Average Time per Assignment</th>
<th>Effective Hourly Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Alignment - Trusted Workers - Dev 0.7</td>
<td>July 02, 2013</td>
<td>1,995 / 2,000</td>
<td>15 days</td>
<td>(TODAY)</td>
<td>1 minute 58 seconds</td>
<td>$4.576</td>
</tr>
</tbody>
</table>
I tried one of his tasks to see, I gave it up at 4 minutes in and about 2/3 of the way through. For the whole hit, I'd have taken about 6 minutes. 10 hits an hour - $1.70 an hour. Restricted to U.S. residents.

This is far too low to be considered a fair wage for a U.S. resident. My performance may be very far off from what others can do. Perhaps I took 4 times or more as long as an average worker would.

My complaint is that any U.S. requester knows what wage rate is required for a U.S. resident to survive. We may not agree on an exact number. But as they say, I know a fair wage when I see it, and this is not it.

Mturk is actually much smaller than what it can appear to be. Something close to requester monopoly has the power to keep wages low. Requester co-operation, explicit or implicit, reinforces this.

Chris Callison-Burch is not unaware, I think, of the mechanics of the wage structure of Mturk.
WORKERS OF THE WORLD UNITE!
For those of you who know Arabic, this is a very solid requester with a very fair pay. Highly recommended for those who want to make some real money. Payment usually take one week depending on the HITs you are doing don't waste your time trying to submit machine translated crap, or random answers to multiple choice questions, you will get blocked instantly.

-----------
Jul 27 2013  |  hala...@h...  | flag | comment

Chris is one of the better requesters on MTurk, if you meet his qualifications and actually do the work as he requires. Glad to see that someone out there can finally work on those Arabic translation HITs that we've all seen for months now.

Jul 27 2013  |  baudelai...@m...  |

Good requester. Everything approved in a couple of days. I had no problems. This is a safe requester to work for.
## qualitative v quantitative

<table>
<thead>
<tr>
<th>TurkOpticon's qualitative attributes</th>
<th>CrowdWorker's quantitative equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>promptness:</strong> How promptly has this requester approved your work and paid?</td>
<td><strong>Expected time to payment:</strong> On average, how much time elapses between submitting work to this Requester and receiving payment?</td>
</tr>
<tr>
<td><strong>generosity:</strong> How well has this requester paid for the amount of time their HITs take?</td>
<td><strong>Average hourly rate:</strong> What is the average hourly rate that other Turker make when they do this requester's HITs?</td>
</tr>
<tr>
<td><strong>fairness:</strong> How fair has this requester been in approving or rejecting your work?</td>
<td><strong>Approval/rejection rates:</strong> What percent of assignments does this Requester approve? What percent of first-time Workers get any work rejected?</td>
</tr>
<tr>
<td><strong>communicativeness:</strong> How responsive has this requester been to communications or concerns you have raised?</td>
<td><strong>Reasons for rejection:</strong> Archive of all of the reasons for Workers being rejected or blocked by this Requester.</td>
</tr>
</tbody>
</table>
Ethics

• Fair pay for workers
• Legal implications of sharing economy
• Ethics of companies like Uber
• Guidelines for human subjects research
Classification System for Human Computation

- Motivation
- Quality Control
- Aggregation
- Human Skill
- Process Order
- Task-request Cardinality
Motivation

How can we motivate people to participate? Even with a low barrier to entry (anyone with an computer can contribute) we still need to make a case why they should contribute.

- Pay
- Altruism
- Reputation
- Enjoyment
- Implicit work
Quality Control

- Reputation systems
- Redundancy and agreement
- Gold standards
- 2nd pass reviewing
- Statistical models
- Defensive task design
- Economic incentives
Aggregation

- Wisdom of Crowds
- Voting
- Prediction markets
- Collection
- Search
- Iterative improvement
- Machine learning
Iowa Electronic Markets: Predictive Accuracy Through Time

Average absolute error in predicting two-party vote shares, 1988-2000

Source: Author's calculations based on data available at: www.biz.uiowa.edu/iem/
Human skill

- Visual recognition
- Language understanding
- Translation
- Reasoning
- Creativity
<table>
<thead>
<tr>
<th>Avoiding dieting to prevent from flu</th>
<th>abstention from dieting in order to avoid Flu</th>
<th>Abstain from decrease eating in order to escape from flu</th>
<th>In order to be safer from flu quit dieting</th>
</tr>
</thead>
<tbody>
<tr>
<td>This research of American scientists came in front after experimenting on mice.</td>
<td>This research from the American Scientists have come up after the experiments on rats.</td>
<td>This research of American scientists was shown after many experiments on mouses.</td>
<td>According to the American Scientist this research has come out after much experimentations on rats.</td>
</tr>
<tr>
<td>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</td>
<td>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</td>
<td>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</td>
<td>Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</td>
</tr>
<tr>
<td>research has proven this old myth wrong that its better to fast during fever.</td>
<td>Research disproved the old axiom that &quot;It is better to fast during fever&quot;</td>
<td>The research proved this old talk that decrease eating is useful in fever.</td>
<td>This Research has proved the very old saying wrong that it is good to starve while in fever.</td>
</tr>
<tr>
<td>Avoiding dieting to prevent from flu</td>
<td>abstention from dieting in order to avoid Flu</td>
<td>Abstain from decrease eating in order to escape from flue</td>
<td>In order to be safer from flu quit dieting</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------------</td>
<td>------------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>This research of American scientists came in front after experimenting on mice.</td>
<td>This research from the American Scientists have come up after the experiments on rats.</td>
<td>This research of American scientists was shown after many experiments on mouses.</td>
<td>According to the American Scientist this research has come out after much experimentations on rats.</td>
</tr>
<tr>
<td>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</td>
<td>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</td>
<td>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</td>
<td>Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</td>
</tr>
<tr>
<td>research has proven this old myth wrong that its better to fast during fever.</td>
<td>Research disproved the old axiom that &quot;It is better to fast during fever&quot;</td>
<td>The research proved this old talk that decrease eating is useful in fever.</td>
<td>This Research has proved the very old saying wrong that it is good to starve while in fever.</td>
</tr>
<tr>
<td>Avoiding dieting to prevent from flu</td>
<td>abstention from dieting in order to avoid Flu</td>
<td>Abstain from decrease eating in order to escape from flue</td>
<td>In order to be safer from flu quit dieting</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>This research of American scientists came in front after experimenting on mice.</td>
<td>This research from the American Scientists have come up after the experiments on rats.</td>
<td>This research of American scientists was shown after many experiments on mouses.</td>
<td>According to the American Scientist this research has come out after much experimentations on rats.</td>
</tr>
<tr>
<td>Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.</td>
<td>in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.</td>
<td>It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.</td>
<td>Experimentaitions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus.</td>
</tr>
<tr>
<td>research has proven this old myth wrong that its better to fast during fever.</td>
<td>Research disproved the old axiom that &quot;It is better to fast during fever&quot;</td>
<td>The research proved this old talk that decrease eating is useful in fever.</td>
<td>This Research has proved the very old saying wrong that it is good to starve while in fever.</td>
</tr>
</tbody>
</table>
New Programming Languages Concepts

Make me a sandwich.

What? Make it yourself.

Sudo make me a sandwich.

Okay.
TurKit: A programming language for the crowd

ideas = []
for (var i = 0; i < 5; i++) {
    idea = mturk.prompt(
        "What’s fun to see in New York City? Ideas so far: " + ideas.join("", "")
    )
    ideas.push(idea)
}
ideas.sort(function (a, b) {
    v = mturk.vote("Which is better?", [a, b])
    return v == a ? -1 : 1
})
New Programming Languages Concepts

- Latency
- Cost
- Parallelization
- Non-determinism
- Iterative improvement
New keyword \textbf{once}

- Costly operations can be marked in a TurKit program with keyword \textbf{once}
- \textbf{once} denotes that an operation should only be executed once across all runs of a program
Quicksort on MTurk

```python
corepare(a, b)
    hitId ← once createHIT(...a...b...)
result ← once getHITResult(hitId)
return (result says a < b)
```

- Subsequent runs of the program will check the database before performing these operations
Quicksort for kittens
When should you mark a function with `once`?

- **High cost** - This is its main usage. Whenever a fn is high-cost in terms of money or time, `once` saves the day.
When should you mark a function with `once`?

- **Non-determinism** - storing results in DB assumes that the program executes in a deterministic way
Wizard of Oz in HCl
Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split base on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.
Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.
Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split base on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.
Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering is not fine enough, leaving heterogeneous clusters that must be edited one line at a time.

One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.
The Human Macro

The Human Macro

Title

Find Creative Commons figure for paragraph

Create Task for Every:

Paragraph

Instructions (with Example)

I need a creative commons licensed image to describe the granite mountain that looks like it was sheared off.

Mechanical Turk Worker Preview

Advertisement

Find Creative Commons figure for paragraph

Instructions

I need a creative commons licensed image to describe the granite mountain that looks like it was sheared off.

Here is the text:

When I first visited Yosemite State Park in California, the rocks were big, the trees were big, the animals were big, the granite mountain that looks like it was sheared off was big.
## Human Macro Examples

<table>
<thead>
<tr>
<th>Request</th>
<th>“Pick out keywords from the paragraph like Yosemite, rock, half dome, park. Go to a site which has CC licensed images [...]”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>When I first visited Yosemite State Park in California, I was a boy. I was amazed by how big everything was [...]</td>
</tr>
<tr>
<td>Output</td>
<td><img src="image-url" alt="Yosemite Park Image" /></td>
</tr>
</tbody>
</table>
VizWiz: Answers to Visual Questions for Blind Users

![Image 1](20-dollar-bill.png) What denomination is this bill?

- 20
- 20

(24s) 20
(29s) 20

![Image 2](parking-lot.png) Do you see picnic tables across the parking lot?

- no
- no

(13s) no
(46s) no

![Image 3](oven.png) What temperature is my oven set to?

- it looks like 425 degrees but the image is difficult to see.
- 400
- 450

(69s) it looks like 425 degrees but the image is difficult to see.
(84s) 400
(122s) 450

![Image 4](chickpeas.png) Can you please tell me what this can is?

- chickpeas.
- beans
- Goya Beans

(183s) chickpeas.
(514s) beans
(552s) Goya Beans

![Image 5](drink.png) What kind of drink does this can hold?

- Energy
- no can in the picture
- energy drink

(91s) Energy
(99s) no can in the picture
(247s) energy drink
Know when work is imminent

<table>
<thead>
<tr>
<th>Time</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>61 seconds</td>
<td>Start app, take picture</td>
</tr>
<tr>
<td>71 seconds</td>
<td>Record the question</td>
</tr>
<tr>
<td>78 seconds</td>
<td>Press send</td>
</tr>
<tr>
<td>221 seconds</td>
<td>Wait for response</td>
</tr>
</tbody>
</table>
Maintain a work pool

- TurKit also experimented with maintaining a group of workers, even when there was no work.
- Created dummy assignments from past assignments, to ensure work.
- When a new request arrived a dummy was replaced with the real request.
- Can be costly to constantly maintain a pool.
Retainer model

- Alternate to maintaining worker pool with dummy tasks
- Hire crowd workers in advance, and pay them a small amount to wait for work to come online
- All them to pursue other work while waiting
- Alert them when our task is ready with a popup box, and pay them for that work too
Improving 10 minute retainer response time

Design
- Baseline
- Alert
- Game
- Reward

Table 1. A tabular representation of Figure 1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Median Completion Time (s)</th>
<th>Baseline</th>
<th>Alert</th>
<th>Game</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retainer</td>
<td>3.46</td>
<td>3.20</td>
<td>3.07</td>
<td>2.94</td>
<td>2.44</td>
</tr>
<tr>
<td>No Retainer</td>
<td>3.75</td>
<td>3.66</td>
<td>3.60</td>
<td>3.50</td>
<td>2.90</td>
</tr>
</tbody>
</table>

Table 2. A tabular representation of Figure 2.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Median Completion Time (s)</th>
<th>Baseline</th>
<th>Alert</th>
<th>Game</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retainer</td>
<td>3.34</td>
<td>3.15</td>
<td>3.10</td>
<td>3.00</td>
<td>2.50</td>
</tr>
<tr>
<td>No Retainer</td>
<td>3.60</td>
<td>3.50</td>
<td>3.40</td>
<td>3.30</td>
<td>2.70</td>
</tr>
</tbody>
</table>

Figure 3. A small reward for fast response (red) led workers to move more quickly than the baseline (blue).

Figure 2. For retainer times under 30 minutes, the system should place a lower cost on the task completion time.

Discussion
- The cost of the retainer model is attractive because it can be used to improve response times without significantly increasing worker stress.
- The game was designed to be engaging and enjoyable, but had a small effect on workers' attention and productivity.
- The alert was designed to be early and short, allowing workers to adjust their attention and focus on the task at hand.

In the alert condition, 75.1% of workers responded within two seconds and 87.2% within two minutes, compared to 64.2% and 50% in the baseline condition.

Figure 1. A graph showing the cumulative distribution of response times for different conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Median Completion Time (s)</th>
<th>Baseline</th>
<th>Alert</th>
<th>Game</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retainer</td>
<td>3.46</td>
<td>3.20</td>
<td>3.07</td>
<td>2.94</td>
<td>2.44</td>
</tr>
<tr>
<td>No Retainer</td>
<td>3.75</td>
<td>3.66</td>
<td>3.60</td>
<td>3.50</td>
<td>2.90</td>
</tr>
</tbody>
</table>

Figure 3. A small reward for fast response (red) led workers to move more quickly than the baseline (blue).

Figure 2. For retainer times under 30 minutes, the system should place a lower cost on the task completion time.

Discussion
- The cost of the retainer model is attractive because it can be used to improve response times without significantly increasing worker stress.
- The game was designed to be engaging and enjoyable, but had a small effect on workers' attention and productivity.
- The alert was designed to be early and short, allowing workers to adjust their attention and focus on the task at hand.

In the alert condition, 75.1% of workers responded within two seconds and 87.2% within two minutes, compared to 64.2% and 50% in the baseline condition.
Studying Economic Markets
Financial Incentives and the “Performance of Crowds”

• Experiment with economic incentives on Amazon Mechanical Turk

• Does compensation change the quantity of work performed (output)?

• Does it change the quality of the work (accuracy)?
Number of tasks done:

- The number of tasks completed increases with increasing pay per task.
- There is a significant positive relationship between pay and the number of tasks completed, as shown by the upward trend in the graph.
- The effect is most pronounced at the lower pay rates, with a gradual increase as pay increases.
- The data suggests that participants are more likely to complete tasks at higher payment levels, indicating a strong dependence of task quantity on extrinsic motivation.

Graphically, the plot illustrates that the number of tasks completed per pay rate exhibits a pronounced upward trend, with the highest number of tasks completed at the highest pay rate of $0.10 per task. The error bars indicate variability across different groups or conditions, and the graph captures the expected increase in task completion with increasing compensation.
As Figure 3 indicates, however, increasing compensation did not improve accuracy, which we measured in two ways: first, using a simple one confirmed quantitatively what is visually apparent in Figure 3 in 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 order across all difficulty levels). The strong and significant dependence of intrinsic motivation (e.g. enjoyment of the task) would have overwhelmed the effect of changes in extrinsic motivation (payment), which can vary by at most $0.10 per task. The strong and significant dependence of accuracy therefore suggests that the range of wage rates studied ought to be sufficient to observe variability in the quality of performance as well.

Nevertheless, the finding is reassuring since, as noted above, one proportionately more of the participants paid $0.10 sorted fewer than 10 sets. These results, in other words, are completely consistent with wage rate differences between the actual and correct ordering, the model is defined as Spearman's rank correlation (\( \rho \)), also known as a multi...
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 and $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).

4.1.1 Design

For each puzzle, we provided a list of words that might be found in the puzzle, although only a subset of the list was actually hidden in the word puzzle. As before, this task allowed us to measure quantity (number of puzzles completed) and quality (fraction of words found per puzzle) independently; but because participants did not know how many words from the list could be found in the puzzle, we could measure the effect of the actual pay on perceived value of the task. Figure 4. Post-hoc survey shows perceived value of the task increases with the actual pay, but is always slightly greater than the actual pay received.
MTurk for social science research

• Many social science experiments require recruitment of a large number of subjects

• MTurk contains the major elements required to conduct research:
  • A participant compensation system
  • A large pool of potential participants
  • A streamlined process for study design, participant recruitment, and data collection
How Do MTurk Samples Compare With Other Samples?

- MTurk population is more diverse than college students (or non-students who reside in college towns)
- Good gender splits
- Good minority representation
- Large number of non-US participants
Active versus Passive Crowdsourcing

• In the first half of the semester we mainly looked at active crowdsourcing, where we explicitly solicit help from the crowd.

• Many applications of crowdsourcing rely on passive information collection from multitudes of individual.
The Best Questions on a First Date

- You would like to learn about your date, some important things that you would like to know are awkward to ask directly.

- Find questions that correlate with what you want to know, but which people are more free about answering publicly.
Are you looking for a partner to have children with?

- Yes
- No

Answer this question privately.
% of long term couples that agree on all 3 answers

Q: Is God important in your life?
Q: Is sex the most important part of a relationship?
Q: Does smoking disgust you?

Q: Wouldn't it be fun to chuck it all and go live on a sailboat?
Q: Do you like horror movies?
Q: Have you ever traveled around another country alone?
What can you do with Crowdsourcing?

- Crowdsourcing is a transformative idea for business and research
- You all are exhibiting hugely creative thinking about it with your final projects
- I am looking forward to seeing what you come up with for the final, and beyond!
Final project details

- Wednesday, May 8th from noon-2pm in Wu and Chen Auditorium (Levine 101)
- 5-7 minutes video for each team, plus 2 minute Q&A
- You must provide links to your at least 1 hour before the presentations begin, and validate that they work.
- Final reports due on the 8th. Submit them before 9am.
Internship opportunities

• I am looking for 2-3 undergraduate researcher assistants to work with me on Crowdsourcing

• Paid summer internships in my lab

• Good experience if you’re thinking about applying to grad schools

• Email me if you’re interested: ccb@upenn.edu
Thanks!