Getting Crowds to Work

Leah Birch
Naor Brown

October 24, 2012
1. Take out a sheet of paper and pen.
2. Write your name at the top.
3. The quiz has 3 questions, so write down 1, 2, 3a, and 3b.
4. The quiz is graded.
1. What was the total number of pages you were required to read for class today?
2. How many acres is Harvard Yard?
3. Estimate the number of calories in the $\frac{1}{2}$ cup of corn. Estimate the number of calories in the barbeque chicken.
Crowdsourcing

Definition

Crowdsourcing is the division and assignments of tasks to large distributed groups of people, both online and offline.
Why Do Incentives Matter?

To find participants
To attract the “right” crowds
Limit cheating
Why Do Incentives Matter?

- To find participants
Why Do Incentives Matter?

- To find participants
- To attract the "right" crowds
Why Do Incentives Matter?

- To find participants
- To attract the "right" crowds
- Limit cheating
The Experiment

Time Period: June 2, 2009 - September 23, 2009

Workers were from Amazon's Mechanical Turk (2159 subjects)

They answered 6 content analysis questions of Kiva.org

Participants randomly assigned 1 of 14 different incentive schemes (financial, social, and hybrid)

Control: Simply offered payment for answering the questions

Additional Information: Demographics were recorded
The Experiment

- Time Period: June 2, 2009 - September 23, 2009
- Workers were from Amazon’s Mechanical Turk (2159 subjects)
- They answered 6 content analysis questions of Kiva.org
- Participants randomly assigned 1 of 14 different incentive schemes (financial, social, and hybrid)
- Control: Simply offered payment for answering the questions
- Additional Information: Demographics were recorded
The Incentive Schemes

1. Tournament Scoring
2. Cheap Talk (Surveillance)
3. Cheap Talk (Normative)
4. Solidarity
5. Humanization
6. Trust
7. Normative Priming Questions
8. Reward Accuracy
9. Reward Agreement
10. Punishment Agreement
11. Punishment Accuracy
12. Promise of Future Work
13. Bayesian Truth Serum
14. Betting on Results
The Results

Figure 1. Performance Distributions - All Conditions

Percent Correct (per question)

Condition

Betting
Norm prime
BTS
Trust
Humanization
Promise opport.
Solidarity
Punish agmt
Punish accuracy
Reward agmt
Reward accuracy
Cheap norm. incentive
Cheap surveillance
Tournament
Control

Figure 2. Performance Distribution - All Conditions

Number of workers

Correct answers

mean = 2.38
The Results

- Subjects performed better than chance, so they actually did attempt to answer the questions.
- Demographic: Users with poor web skills and residence in India performed worst.
- Social incentives did not cause the performance to differ from the control.
- Financial incentives had more correct answers.
- Punishment of disagreement and the Bayesian Truth Serum out performed other Incentive Schemes.
Why Bayesian Truth Serum?

The condition confused workers. Though confused, they had to think about how others responded. Thus, workers were more engaged for BTS, yielding better performance.
The condition confused workers.
Why Bayesian Truth Serum?

1. The condition confused workers.
2. Though confused, they had to think about how others responded.
Why Bayesian Truth Serum?

1. The condition confused workers.
2. Though confused, they had to think about how others responded.

Thus, workers were more engaged for BTS, yielding better performance.
Why Punishment of Disagreement?
Amazon’s Mechanical Turk can ban workers if requesters reject their work.
Amazon’s Mechanical Turk can ban workers if requesters reject their work.

The wording of the condition made workers overly cautious, and so more likely to answer to the best of their ability.
Problems with Experiment?

1. Too many conditions being tested.
2. The incentive structures were not followed exactly. (Everyone was paid the same amount in the end.)
Problems with Experiment?

1. Too many conditions being tested.

2. The incentive structures were not followed exactly. (Everyone was paid the same amount in the end.)
Limitations of Crowdsourcing

- There is a difficulty in obtaining groups of people with specific skills or knowledge.
- Experts would have to be enticed with targeted rewards.

How do we reconcile these issues?
Definition

The division and assignment of tasks to targeted crowds with specific knowledge and specialized skills.
Potential of Communitysourcing

1. Can communitysourcing successfully enlist new users groups in crowd work?
2. Can communitysourcing outperform existing crowd sourcing methods for expert, domain-specific tasks?
3. How does communitysourcing compare to traditional forms of labor, in terms of quality and cost?
Umati: The communitysourcing vending machine
Design Considerations

1. Problem selection: what problem are we trying to solve?
2. Task selection: short-duration, high-volume tasks that require specialized knowledge or skills specific to a community but widely available within that community
3. Location selection: targeting the right crowd, repelling the wrong crowd
4. Reward selection: something that the community finds interesting and valuable
5. Context selection: preferably during idle time (where there is "cognitive surplus")
Communitysourcing in action

Can you think of some communitysourcing projects? Example: a slot machine in an airport with travel review tasks.

1. Problem?
2. Task?
3. Location?
4. Reward?
5. Context?
Communitysourcing in action: Umati

1. Problem: Grading exams is a painful, high-volume task
2. Task: Grading the exam
3. Location: The CS department building, in front of lecture hall
4. Reward: Food!
5. Context: Plenty of boring waiting time before (and probably during) lecture
How Umati works: The vending machine
How Umati works: Touch screen scenarios
How Umati works: The grading interface
How Umati works: Spam detection

1. Low-quality responses are a common issue in crowdsourcing.

2. Umati provides a disincentive from gaming the system.

3. Single-spam detection system; some tasks have a known correct response, and if 2 of these are incorrectly filled by the participant, then the participant is logged off and blacklisted.

Can you think of any other methods to disincentivize people from spamming?
Design of the Study

3 groups of participants graded exams in introductory class, CS2:

1. Umatiers
2. Turkers
3. Experts (CS2 teaching staff)

The gold standard tasks had known answers given by the professors themselves.
Grading on Umati  Results

1. Umati was filled with $200 of candy, and refilled twice during the week.
2. Umatiers were offered one credit per answer, where most snacks were priced at 20 credits.
3. Umatiers were essentially paid 2.25 cents per answer in candy, and answers took about 15 seconds per question.
4. Umatiers would spend 5 minutes maximum at a time.
5. 328 participants graded 7771 exam answers in 1 week (even though the machine ran out of candy for 3 days).
6. 61 of the 328 were blacklisted for failing gold standard tasks, probably because of the novelty of Umati.
7. 81% of Umatiers majored in EE or CS (success in communitiesourcing!).
Grading on Mechanical Turk Results

1. Attempt of “communitysourcing” MTurk by having Turkers take a qualification exam before being allowed to participate.

2. Those who passed the qualification test were paid 2.8 cents per answer, and those who did not pass were paid 2.6 cents per answer.

3. 1050 grades in 3 days, with 46 unique Turkers.

4. 16 of them failed the spam detection.

5. Conclusion: a qualification exam is not enough to attract qualified Turkers (failure in communitysourcing!).
1. 10 staff were paid 34 cents per answer.
2. All experts passed the gold standard test.
3. Median scores were used as a “single” expert.
Interesting Take-aways

1. Umatiers, on aggregate, grade similarly to experts.
2. Turkers do not grade very accurately.
3. Individual experts and Umatiers agree strongly with all experts on assigning full credit or no credit.
4. Single experts have lower variance than Umatiers on assigning partial credit.
5. Price-accuracy tradeoff: If we keep cost constant, Umati is more accurate than a single expert grader. For 34 cents, we can buy one expert with 78.3% accuracy, or 15 Umatiers with 8.3% accuracy.
6. Therefore, Umati was the cheapest and most accurate way to grade this exam.
Interesting Take-aways

Figure 8. Agreement between conditions and the median of the 10 expert graders. Umati users’ response distributions were much closer to the experts than distributions of Mechanical Turk users.
Interesting Observations

1. People loved using Umati, and sometimes there would be a long line.
2. Multiple people would use Umati at the same time, and people would argue over answers.
3. Some users were unjustly blacklisted.
4. One group was really hungry and just tried prying open the door.
Limitations of Umati study

What are some limitations of Umati?

1. A student can cheat by self-grading (even though it is difficult).
2. Study was over 1 week; would it still sustain its charm over a long time period?
3. It was not compared to expert platforms such as oDest or Stack Overflow.
Limitations of Umati study

What are some limitations of Umati?

1. A student can cheat by self-grading (even though it is difficult).
2. Study was over 1 week; would it still sustain its charm over a long time period?
3. It was not compared to expert platforms such as oDest or Stack Overflow.
What improvements would you suggest for making Umati better?
What improvements would you suggest for making Umati better?

1. To prevent possible cheating, grade tests that are from a different school.

2. Lower-stake grading, such as for homework or practice exams.

3. Higher-value rewards, such as video game currency instead of food.
Waze
**Plate-mate**

![Image of a plate with food]

<table>
<thead>
<tr>
<th>Dinner</th>
<th>kcal</th>
<th>fat (g)</th>
<th>carb (g)</th>
<th>protein (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Yellow Corn (0.50 cup)</strong></td>
<td>500</td>
<td>3.9</td>
<td>61.6</td>
<td>7.0</td>
</tr>
<tr>
<td><strong>Barbeque chicken breast</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Chicken Breast Meat and Skin (Broilers or Fryers) (1.00 breast, bone removed)</strong></td>
<td>249</td>
<td>10.4</td>
<td>0</td>
<td>30.2</td>
</tr>
<tr>
<td><strong>Barbeque Sauce (Low Sodium, Canned) (0.14 cup)</strong></td>
<td>26.6</td>
<td>0.8</td>
<td>4.5</td>
<td>0.6</td>
</tr>
<tr>
<td><strong>Beef Steak (0.92 medium steak (yield after cooking, bone removed))</strong></td>
<td>471.3</td>
<td>28.1</td>
<td>0</td>
<td>51.0</td>
</tr>
<tr>
<td><strong>Hominy (White, Canned) (0.44 cup)</strong></td>
<td>52.8</td>
<td>0.8</td>
<td>10.4</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>Ketchup (2.00 tbsp)</strong></td>
<td>30</td>
<td>0.1</td>
<td>7.5</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Beef Steak (0.06 medium steak (yield after cooking, bone removed))</strong></td>
<td>440.5</td>
<td>26.2</td>
<td>0</td>
<td>47.7</td>
</tr>
</tbody>
</table>

*Figure 1.* The PlateMate user interface. Users upload photographs of their meals, which are processed through Mechanical Turk to produce a list of foods, serving sizes, and nutrition information.
Galaxy Zoo

Few have witnessed what you're about to see
Experience a privileged glimpse of the distant universe, observed by the Sloan Digital Sky Survey and Hubble Space Telescope

Classify Galaxies
To understand how galaxies formed, we need your help to classify them according to their shapes. If you're quick, you may even be the first person to see the galaxies you're asked to classify.

Begin Classifying

How Do Galaxies Form?
Figure 4. Find-Fix-Verify identifies patches in need of editing, recruits workers to fix the patches, and votes to approve work.
Duolingo

With Duolingo you learn a language for free while helping to translate the web

Sign up for Duolingo

Facebook Connect to Facebook Twitter Connect to Twitter

or Sign up with email
99Designs

Logo design, web design & more!
Find out why we’re the fastest growing design marketplace in the world

See examples
- Logo design
- Business card design
- T-shirt design
- Postcard or flyer design
- Brochure design
- Product packaging design
- Web page design
- Mobile app design
- Icon or button design
- Banner ad design
- Illustration
- Book cover design

Get started now
Wikipedia
Facebook

Vote for Translations

Displaying 10 phrases in Afrikaans.

Translate Where do the studies take place?
No description for this phrase is available.

Waar het die studies plaas gevind?
Votes:

Translate Entertainment

advertiser category
Pop Quiz Answers

1. Q: What was the total number of pages you were required to read for class today?
   A: 20 pages

2. Q: How many acres is Harvard Yard?
   A: 25 acres

3. Q: Estimate the number of calories in the \( \frac{1}{2} \) cup of corn.
   A: 303 calories

   Q: Estimate the number of calories in the barbeque chicken.
   A: 275.6 calories