Handle with Care: Opportunities and Threats of Amazon Mechanical Turk Experimentation

Gabriele Paolacci
Erasmus University Rotterdam

Joint work with:

Jesse Chandler (PRIME Research & University of Michigan)
Pam Mueller (Princeton University)
1) What? Microtasks (e.g., image tagging)
2) For how much? LOW wages (as low as $1 / hour)
3) Who? Mostly US and India, avg. age 34, 60% female
4) Why? Mix of intrinsic and extrinsic motivations
Mechanical Turk and Behavioral Research

• Survey on SPSP and SJDM mailing lists (N = 369)
  • 50% have used Mturk. Because of its:
    • Speed (58%)
    • Affordability (31%)
    • Diversity (31%)

• How is data quality?
  • JDM classics do replicate: http://experimentalturk.wordpress.com
  • Similar to to other pools (Paolacci, Chandler, & Ipeirotis, 2010)
  • Not affected by pay rate (Mason & Watts, 2011)

Mturk is good.
Are we?
A neglected concern: Participant Non-Naivety

• Foreknowledge is often bad

• Two potential causes of non-naivety:
  1. Crosstalk
  2. Repeated participants

• Exacerbated problem on Mturk:
  • At least half of us share this subject pool
  • It is not continuously replenished (about 500,000 workers)

• Only 4% researchers seem concerned
An example

A worker on a forum:

It makes me wonder if some of the folks posting surveys are somehow unaware that turkers take hundreds of surveys. If they're unaware, they don't know what they're doing.

If they are aware, why do I see so many surveys cutting-and-pasting questions like, as an example, these:

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

These questions can only work to measure what they're supposed to measure the very first time a survey-taker sees them. After that, the cat's out of the bag. Why would a serious researcher put that sort of content into a survey that goes to a pool of professional survey takers? It's a tainted pool. Do they not notice that MTurkers get those questions right way more than the general population? Again, I suspect many of these researchers simply don't care. They're either students, or they're just ripping off whomever gave them the money for the research.
Mturk Crosstalk

- Mturk forums:
  - turkernation.com
  - mturkforums.com
  - reddit.com/r/mturk

- On these forums turkers:
  - talk about HITs
  - talk about Requesters
  - link to HITs

---

Not really funny but a good heads-up. A new thing I have seen in a couple of surveys today is special instructions about only copying part of the “code” at the end. For example, the last one I did said this:

On the next page, you will see a complete code which will consist of three letters, two numbers, and then the letter "x" followed by three additional numbers. In order to receive credit on Mechanical Turk, copy only the three letters and two numbers that precede the "x". If you copy the three numbers that follow the "x", we will assume that you did not follow our directions and you will not be paid.

First they came for the Communists, and I didn’t speak up, because I wasn’t a Communist. Then they came for the Jews, and I didn’t speak up, because I wasn’t a Jew. Then they came for the Catholics, and I didn’t speak up, because I was a Protestant. Then they came for me, and by that time there was no one left to speak up for me.
Is crosstalk a problem?

Worker survey (N = 320):
• 26% reported knowing someone else who used MTurk personally
• 28% reported reading forums and blogs about MTurk
• 13% reported ever seeing a discussion about the contents of a study

<table>
<thead>
<tr>
<th>Ranked Topics of conversation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How much the HIT pays</td>
</tr>
<tr>
<td>2. Requester reputation</td>
</tr>
<tr>
<td>3. How long the HIT takes</td>
</tr>
<tr>
<td>4. How difficult it is to complete</td>
</tr>
<tr>
<td>5. How to successfully complete the HIT</td>
</tr>
<tr>
<td>6. How fun the HIT was</td>
</tr>
<tr>
<td>7. Purpose of the HIT</td>
</tr>
</tbody>
</table>
Is crosstalk a problem?

• Answer: Probably not a huge one. Still:
  • Frequently monitor forums
  • Ask people how they reached your HIT
  • Care about your reputation!
Repeated Participants

• How likely is it that:
  • *any MTurk worker has done a similar HIT before?*
  • *a worker in my HIT has done something similar before?*

• Method:
  • We pooled past HITs from us and other researchers:
    • 132 batches (studies)
    • 16,408 HITs (individual surveys taken)
  • We looked at concentration of working activity
  • We supplemented this with our Worker survey data
Superturkers

10% of HITs

41% of HITs

Number of HITs Completed

Activity Level of Worker

99th percentile
90th percentile
0-89th percentile

Gabriele Paolacci – Erasmus University Rotterdam
Oxford, 14/01/2013
Superturkers

The ULTIMATE SUPERTURKER

Gabriele Paolacci – Erasmus University Rotterdam

Oxford, 14/01/2013
Classics should better be avoided

Common Paradigms on MTurk

![Bar chart showing percentage of workers who previously participated in different types of studies.](chart.png)
And common manipulations too.
They know who you are!

turkopticon.differenceengines.com

<table>
<thead>
<tr>
<th>HIT Group »</th>
<th>FAIR</th>
<th>FAST</th>
<th>PAY</th>
<th>COMM</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabriele Paolacci A1T8B22QTO7H8N</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>3 / 5</td>
<td>NO DATA</td>
<td>On July 12, I did &quot;Complete a 3-minute survey&quot; for $0.20. Approved and paid within a few days. Jul 19 2012</td>
</tr>
<tr>
<td>Gabriele Paolacci A1T8B22QTO7H8N</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>good Jul 29 2012</td>
</tr>
<tr>
<td>Gabriele Paolacci A1T8B22QTO7H8N</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>Accidentally submitted without code and he responded right away and paid. Good guy. Sep 03 2012</td>
</tr>
<tr>
<td>Gabriele Paolacci A1T8B22QTO7H8N</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>&quot;Make some hypothetical choices in a max 3-minute survey&quot;, quick, easy, and paid a few hours later. Sep 03 2012</td>
</tr>
</tbody>
</table>
You might have a loyal following

![Bar chart showing data on reading forums and following requesters]

- Read forums
- Follow requesters
- Follow academic requesters

turkalert.com

Gabriele Paolacci – Erasmus University Rotterdam

Oxford, 14/01/2013
Preventing Repeated Participants

- Block previous workers

- “Please don’t do this survey if you’ve done it already”.

- Direct turkers to same page, change the redirect

- “Have you done anything similar before?”

- Qualifications (duplicates won’t see your HIT)

- Use our Qualtrics/MTurk script (this afternoon)
Alternatives?

- Crowdflower.com
- Socialsci.com
- Survata.com
- Surveysignal.com
- ...

Oxford, 14/01/2013
Advanced Uses of Mechanical Turk

• Not just one-shot surveys...
  • Cross-Cultural Design
  • Longitudinal Design
  • Synchronous Design
Cross-Cultural Design

• Self-reported:
  • Ask workers demographic questions
  • Run experiment
  • Compare ex-post

• Qualifications:
  • Use “country of origin” qualification

• Issues:
  • “500.000 workers from 190 countries” but...
  • Beware of (mturk) Indians...
Longitudinal Design

• Response rates among recontacted participants:
  • 60%+ response rate after few months (e.g. Berinsky et al. 2012)
  • 44% after one year (our data)
    • 75% among 10% most productive workers)

• Self-reported demographics are consistent (e.g. Suri and Mason 2010)
Synchronous Design

Four steps (Mason and Suri 2011):

1. Recruitment of participants into a panel
   • Keep track of Worker IDs

2. Notification of a start time
   • Send email to Workers one day before
   • 11am-5pm ET work as a start time
   • If you aim at N, email 3N

3. Waiting room that accumulates participants up to a threshold

4. Handling attrition
   • Timeout, Default action
   • Discard observations with <x human actions
Conclusions

• There are neglected concerns and opportunities:
  • Participants are not necessarily naive
  • Not just one-shot surveys

http://experimentalturk.wordpress.com