Getting Europe into the Turk

EMAC 2015

Gabriele Paolacci

Rotterdam School of Management
Erasmus University Rotterdam
(Paid) Crowdsourcing

Already five years ago...

22,000,000+ workers

$500,000,000+ profits/year

$1,000,000,000 wages

(Frei, 2010)
Amazon Mechanical Turk (mturk.com)

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
Research on MTurk

• Why Mturk?
  • **Speed** – Hundreds of responses in a few hours
  • **Affordability** – Few cents per participant
  • **Diversity** – Age 18-80, different education levels, etc.

• Data quality meet standards
  • [http://experimentalturk.wordpress.com](http://experimentalturk.wordpress.com) ("my blog")
  • Paolacci, Chandler, & Ipeirotis (2010) *Judgment and Decision Making*
  • Buhrmester, Kwang, & Gosling (2011) *Perspectives in Psychological Sciences*
  • Horton, Rand, & Zeckhauser (2012) *Experimental Economics*
  • Crump, Donnell, Gureckis (2013) *PLoS ONE*
  • **Klein at al. Many Labs**
  • ...
### Worker Perspective

#### HITs

**1-10 of 2848 Results**

| Requester: | CrowdFlower | HIT Expiration Date: | Apr 13, 2013 (2 days 14 hours) | Reward: | $0.20 |
| Time Allotted: | 15 minutes | HITs Available: | 9 |

**Description:** Visit the following link and complete a survey which will help us evaluate a set of possible company names.

| Requester: | CrowdFlower | HIT Expiration Date: | Apr 16, 2013 (6 days 8 hours) | Reward: | $0.45 |
| Time Allotted: | 15 minutes | HITs Available: | 3 |

**Description:** We value your opinions, please answer each question honestly and thoughtfully. Thanks!

| Requester: | care2team | HIT Expiration Date: | Jun 7, 2013 (8 weeks 1 day) | Reward: | $0.01 |
| Time Allotted: | 60 minutes | HITs Available: | 4592 |

**Description:** Given the company/organization name & address, find the URL for their official website

**Keywords:** data, collection, organizations, website, URLs

**Qualifications Required:** Masters has been granted
Some typical questions (and the outline for today)

1. “Can I run studies other than one-shot surveys?”
2. “How much should I pay?”
3. “How to ensure attention in my study?”
4. “Are participants naïve?”

...just ask 😊
1. Advanced Uses of Mechanical Turk

Not just one-shot surveys...

• Cross-Cultural Design

• Longitudinal Design

• Synchronous Design

• (Workers as Research Assistants)

• (Field Experiments)
Cross-Cultural Design

• Self-reported:
  • Ask demographics, run study, compare ex-post

• Qualifications:
  • Use “country of origin” qualification

• Issues:
  • “500,000 workers from 190 countries” but...
  • We know they are fewer
  • Unclear how Amazon accepts international workers
  • Data quality may be inconsistent
Follow-ups, Longitudinal Designs

• You can message Turkers who completed your studies (e.g., tell Wave 1 subjects about the Wave 2 task)
  • using the Web interface (one at a time)
  • using CLT (many at a time; see script on the blog)

• Response rates:
  • **70%+** after a few weeks (e.g., Chandler et al., in press)
  • **60%+** after few months (e.g. Berinsky et al. 2012)
  • **44%** after 1 year (e.g., Chandler, Mueller, & Paolacci 2014)
    (75% among 10% most productive workers)
Synchronous Design

Four steps (Mason and Suri 2011):

1. Recruitment of subjects 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 subjects up to a threshold

4. Handling attrition
   • Timeout, Default action
   • Discard observations with <x human actions
2. “How to ensure attention in my study?”

Myth: Turkers are less attentive than lab subjects.

Fact: Turkers are (at least) as attentive as lab subjects.

• Attention checks (e.g., Paolacci, Chandler, & Ipeirotis 2010)

• Especially 95%+ approval workers (Peer, Acquisti, & Vosgerau 2013)

• Not because they know the checks (Hauser & Schwarz in press)

Really need checks? Use novel ones.

• See Hauser & Schwarz in press (BRM) for a discussion
3. “How much should I pay?”

- The more you pay, the quicker you reach your N

- Payment seems not to affect data quality. But...
  - Studies conducted time ago, e.g.,
    - Buhrmester, Kwang, & Gosling (2011)
    - Mason & Watts (2009)
  - www.wearedynamo.org/
  - Ethical reasons to pay fairly!

- Many expect at least $0.10/minute ($6/hour)

- Academics have a bad reputation. Let’s change this!
Workers know who you are

<table>
<thead>
<tr>
<th>Name</th>
<th>FAIR:</th>
<th>FAST:</th>
<th>PAY:</th>
<th>COMM:</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabriele Paolacci</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</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</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</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>
One more selfish reason to pay fairly (Twitter)
Talking about Twitter...

Some people who frequently tweet about MTurk:

@Rochelle
@TurkerNational
@kbshee
@David_G_Rand
@RolfZwaan
@gpaolacci
4. “Are participants naïve?”

- We often want participants to be naïve to:
  - experimental manipulations
  - measures

- Many reasons to worry on MTurk:
  - Turkers self-select into tasks
  - It is crowded with researchers
  - It is not continuously replenished
  - Turkers can “follow” requesters (turkalert.com)
  - Turkers can discuss studies on forums (e.g., reddit)
  - Also...
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 some survey data
• 7,498 workers completed 16,408 HITs (avg. worker = 2.24 HITs)

Chandler, Mueller, & Paolacci (Behavior Research Methods 2014)
With no exposure to the intermediate option we obtained a significant endowment effect. Participants exposed to the intermediate option show no endowment effect.

Classics are extremely common. Chandler, Mueller, & Paolacci (Behavior Research Methods 2014)
Consequences of Non-Naivety: An example

Found 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.
Non-naivety systematically affects effect sizes

Investigation building on the “Many Labs” replication project
• 1000 turkers completed 12 two-condition JDM experiments
• (72%) Completed them again after some time
• What happens to the effects?

We manipulated:
• W1-W2 delay (few days vs. about a week vs. about a month)
• Visual similarity between W1 and W2 (same vs. different visual)
• Condition (same vs. different)

Chandler, Paolacci, Peer, Mueller, & Ratliff (Psych Science, forthcoming)
Non-naivety may reduce effect sizes

Table 1. Experimental Effects of Condition and Wave

<table>
<thead>
<tr>
<th>Task</th>
<th>Wave 1</th>
<th></th>
<th>Wave 2</th>
<th></th>
<th></th>
<th>Condition × Wave interaction (Wald)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Comparison of conditions</td>
<td>Effect size (Cohen’s $d$)</td>
<td>Comparison of conditions</td>
<td>Effect size (Cohen’s $d$)</td>
<td></td>
</tr>
<tr>
<td>Allow/forbid</td>
<td></td>
<td>$\chi^2(1, N = 638) = 24.61^{***}$ 0.40</td>
<td></td>
<td>$\chi^2(1, N = 638) = 7.32^{***}$ 0.22</td>
<td></td>
<td>4.35*</td>
</tr>
<tr>
<td>Anchoring: births</td>
<td></td>
<td>$\kappa(489) = 25.83^{***}$ 2.11</td>
<td></td>
<td>$\kappa(557) = 20.65^{***}$ 1.65</td>
<td></td>
<td>6.40**</td>
</tr>
<tr>
<td>Anchoring: Mount Everest</td>
<td></td>
<td>$\kappa(608) = 31.39^{***}$ 2.55</td>
<td></td>
<td>$\kappa(559) = 20.83^{***}$ 1.79</td>
<td></td>
<td>37.63***</td>
</tr>
<tr>
<td>Anchoring: Chicago</td>
<td></td>
<td>$\kappa(595) = 18.06^{***}$ 1.46</td>
<td></td>
<td>$\kappa(545) = 12.62^{***}$ 1.02</td>
<td></td>
<td>3.28†</td>
</tr>
<tr>
<td>Anchoring: NY to SF</td>
<td></td>
<td>$\kappa(552) = 12.78^{***}$ 1.04</td>
<td></td>
<td>$\kappa(602) = 10.73^{***}$ 0.86</td>
<td></td>
<td>1.45</td>
</tr>
<tr>
<td>Gain vs. loss framing</td>
<td></td>
<td>$\chi^2(1, N = 638) = 56.15^{***}$ 0.62</td>
<td></td>
<td>$\chi^2(1, N = 638) = 21.55^{***}$ 0.37</td>
<td></td>
<td>2.88†</td>
</tr>
<tr>
<td>Imagined contact</td>
<td></td>
<td>$\kappa(636) = 4.38^{***}$ 0.35</td>
<td></td>
<td>$\kappa(636) = 1.79^\dagger$ 0.14</td>
<td></td>
<td>6.07*</td>
</tr>
<tr>
<td>Low vs. high scales</td>
<td></td>
<td>$\chi^2(1, N = 638) = 27.09^{***}$ 0.42</td>
<td></td>
<td>$\chi^2(1, N = 638) = 47.24^{***}$ 0.57</td>
<td></td>
<td>0.66</td>
</tr>
<tr>
<td>Norm of reciprocity</td>
<td></td>
<td>$\chi^2(1, N = 637) = 8.50^{**}$ 0.23</td>
<td></td>
<td>$\chi^2(1, N = 636) = 0.90$ 0.08</td>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>Quote attribution</td>
<td></td>
<td>$\kappa(633) = 6.13^{***}$ 0.48</td>
<td></td>
<td>$\kappa(636) = 3.56^{***}$ 0.28</td>
<td></td>
<td>4.67*</td>
</tr>
<tr>
<td>Retrospective gambler's fallacy</td>
<td></td>
<td>$\kappa(429) = 5.43^{***}$ 0.43</td>
<td></td>
<td>$\kappa(558) = 2.90^{***}$ 0.23</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>Sunk costs</td>
<td></td>
<td>$\kappa(612) = 5.30^{***}$ 0.42</td>
<td></td>
<td>$\kappa(636) = 2.36^{*}$ 0.19</td>
<td></td>
<td>0.56</td>
</tr>
</tbody>
</table>

Note: Positive $t$ values indicate that the effect was in the theoretically predicted direction. Reported $t$ values and degrees of freedom assume equality of variance unless Levine’s test indicated that this assumption was violated. Fractional degrees of freedom were rounded down. The chi-square tests of the interaction between condition and wave used a generalized estimating equation treating Wave 1 and Wave 2 data as nonindependent.

$^\dagger p < .10. ^* p < .05. ^{**} p < .01. ^{***} p < .001.$

Chandler, Paolacci, Peer, Mueller, & Ratliff (Psych Science, forthcoming)
Preventing Repeated Participants

• You’ve run Study 1, and now you want to run Study 2
• How can you prevent repeated participants?
Preventing Repeated Participants

- Block previous workers
- “Don’t do this survey if you’ve done it already”
- “Have you done anything similar before?”
- Direct turkers to same page, change the redirect
- TurkGate.net, PsiTurk.org, Turkprime.com, uniqueturker.myleott.com, Qualtrics/MTurk script (see blog)
- Qualifications (see blog)
1. Create a Qualification (e.g., “Study 1”); keep in mind that the name will be visible to Workers in MTurk/Manage/Qualification Types.

2. Download from Manage Workers your “WorkerList”, i.e., the list of workers who participated in any of your study in the past. This may take some time. Open the file and create a new sheet (let’s call it “Sheet1”, the default on my Excel).

3. Download from MTurk the Results batch of your “Study 1” (i.e., of the study whose participants you want to exclude). Copy the WorkerIDs of Study 1 participants and paste them in a column of “Sheet1” in the “WorkerList” file (e.g., Column A).

4. In the original sheet of the WorkerList, find the “UPDATE-Study 1” column. This is the column you need to modify in order to add or change an existing Qualification that concerns “Study 1”. In the top empty cell of that column (row 2; A2 is the location of the first worker in the list), write the following formula (assuming “Sheet1” and Column A are what you used in the steps above):
   \[=IF(ISERROR(MATCH(A2,Sheet1!A:A, 0)), "", 1)\]

5. Copy this formula and paste it into the entire “UPDATE-Study 1” column. People who participated in Study 1 have now a “1” in the “UPDATE-Study 1” column of your WorkerList file. You can double-check the success of the entire procedure by summing up the 1s in the column, which should give you the N of Study 1.

6. Make sure the “UPDATE-Study 1” column is made of values and not of formulas (Select the column, Copy, Paste special, Values), and then delete the unnecessary “Sheet1”. Save your file in .csv format.

7. Upload your “WorkerList” file in the Manage Workers page, and confirm that you want to assign the “Study 1” Qualification to a bunch of workers (the amount should again be equal to the N of your Study 1).

8. When you design your Study 2 HIT, require that the Qualification “Study 1” “has not been granted”. People who were assigned that “1” in your WorkerList file (or any other value) will not be able to participate in your study.
What kind of study is suitable for testing on MTurk?

What kind of study is not suitable for testing on MTurk?

- “Common” paradigms if they require naivety
- Representative samples
- Deceptive studies
- Incentivized answers that can be found online

Align expectations and you can run even:

- High attention studies (e.g., Crump et al. 2013)
- Long studies (30+ minutes) (e.g., Mor, Morris, and Joh 2013)
Some questions I got in the recent past

Does a paper need to have studies besides MTurk to be reliable? i.e., will reviewers have issues?

• Published papers with only MTurk data exist
• But having further samples:
  • increase confidence in your effect
  • increase skeptics’ confidence in your MTurk studies
• Be mindful about issues and ready to defend your choices

What if results differ across samples?
• Moderators (e.g., demographics)?
• Issues with either of the pool (e.g., non-naivety)?
Some questions I got in the recent past

Can we rely on self-filtering when we are only interested in certain participants (e.g., females)?

Better to:

1. Ask relevant info to everyone
2. Recontact eligible people later

or

1. Ask relevant info to everyone
2. Redirect eligible people to your study (others to something else)

or

1. Conduct study with everyone
2. Discard non-eligible people
How to generate randomized codes for compensation?

On Qualtrics, you can use the “Piped Text...” function to display:

1. A random number you generated: http://www.qualtrics.com/university/researchsuite/articles/random-number-generator/ or

2. Respondents’ unique ResponseID with this code: ${e://Field/ResponseID}
Thanks!

Make sure you check my MTurk blog:
http://experimentalturk.wordpress.com
(especially Resources section for papers and scripts)