Doing Human Subjects Research

using Online Platforms

Sample Online Platforms

Amazon Mechanical Turk (MTurk) – mturk.com

- Population size: Over 500,000
- Can screen by approval rating, by location or creating custom qualifications (incl. having taken part in previous studies)

CrowdFlower (CF) – crowdflower.com

- Population size: About 10,000
- Cannot screen by approval rating, or previous participation, but can screen by location and language

Prolific Academic (ProA) – prolific.ac

- Population size: About 60,000
- Can screen by approval rating, by demographics and by certain qualifications such as having taken part in previous studies

Problems

with MTurk (and most other platforms)

Problems with Mechanical Turk

Complaints from Researchers:

- Naivety of the Respondents
  - Professional survey takers
  - Completing the same survey more than once
  - Favoring particular researchers (or not) – Turkalert
  - Limited sampling
- Respondents Organizing Among Themselves (Turkopticon and mturk forum)
  - Potentially affects reputation of researcher
  - If deception is involved, workers may disclose deception on Turkopticon
- Limitations of the Platform
  - Participants prohibited from downloading software
  - Participants prohibited from disclosing identifiable information, including e-mails

Problems with Mechanical Turk

Complaints from Workers:

- Insufficient payment
- Lack of response to problems from researchers
- Issues with survey content (e.g., racially or politically motivated)
- Attention check questions are too picky

  Jane saw Ben's sweater in Mary's locker and demanded that she give it back to him. Who is "she" referring to?

Problems with Mechanical Turk

Complaints from the Scientific Community:

- An MTurk sample is not a random sample
  - It is a convenience sample
- MTurk users are not diverse
  - Demographics do not reflect the US population
- MTurk data is not reliable
  - Participants can lie, cheat, or simply not pay attention

Running a study

while avoiding the mentioned problems!

The basics

1. Create a HIT (Human Intelligence Task)
2. Specify workers
3. Make a pre-payment
4. Run the HIT
5. Pay workers who did the HIT right

Create a HIT

- Give participants ample time
  - Tell them not to rush
- Include your email address and/or a "did anything go wrong" field
  - Be available while the HIT is running for troubleshooting
- Make sure your HIT works on any browser/device
  - and/or stipulate requirements carefully
- Using an external survey system: good idea!
  - Generate a random "code" at the end that you save to your DB and that workers use to submit their HIT
  - Use this code to verify worker participation

Specify workers

- Set qualifications!
  - US participants perform better than those from outside the US
  - Get people with a high HIT rate and some number of completed HITs
  - "Masters" are not really better than general workers
- Prevent repeat participation
  - Collect IP addresses (to detect users with multiple accounts; doesn't happen much but still)
  - Run everything as a single batch (repeat participation automatically prevented)
  - If you want to run multiple batches, create a "qualification" to exclude past workers (https://blog.mturk.com/4d0732a7319b)
  - The latter can also be used to encourage repeat participation! (Longitudinal study)

Make a pre-payment

The "platform overhead" on MTurk is 40%

Pay a decent amount ($8/hr is a good guideline)

- Note: MTurk workers tend to be very quick!

You can use bonus to improve effort, or to include an optional part

- Example: 5min screening questionnaire for 50ct, $2.50 bonus for 15min follow up if selected (make sure they can't find out what makes them selected tho!)
- Example: $2.50 bonus to install an app and try it out (forcing app installation is not allowed)

Run the HIT

- Can easily get 100+ participants in a matter of hours!
  - Make sure your server can handle this
  - Be available while the HIT is running for troubleshooting
- Participation slows down near the end
  - This is due to workers "locking up" the HIT

Pay workers who did the HIT right

Filter bad participants:

- Attention checks (to check for reading)
- Open text questions (to check for effort)
- Reverse-coded items (to check for consistency)
- Timers (to check for effort)

In general, I throw out 10-15% of the data

I **only** withhold payment if cheating is indisputable (otherwise it's not worth it)

Pay promptly

Helps with getting a good reputation

Solving problems

What if I get into trouble?

What if workers complain?

- MTurk workers are very persistent, but they can get angry if you wrong them
  - Make sure to always respond promptly and politely
- Make sure you have a good reputation
  - Workers talk among themselves… having a bad rep means workers will avoid your HIT
- In general, if you follow the tips, you should be OK
  - React quickly, pay promptly, don't withhold payment unless cheating is 100% obvious

What if I get flagged by the MTurk platform?

The platform owners (Amazon) don't act unless they receive complaints

If you get reported, your HIT may get removed

- Too many removed HITs and your account will be deleted (and your credit card flagged)
- You have very little recourse when you get removed… The rules are rather unclear

In general, there is very little support from Amazon (not like Qualtrics), so keep that in mind

What if I get flagged by the MTurk platform?

Examples:

- **You are not allowed to ask participants for PII**

  But even when you *fake* collecting PII (and have the IRB to back this up), you can get kicked off

- **You are not allowed to ask participants to login to a service, other than for completing the HIT**

  But even when you make clear that the login is needed for completing the HIT and no extraneous information is collected, you can get kicked off

- **You are not allowed to *require* the participants to download spyware or keyloggers**

  But even when you ask participants to download and install a legitimate app *optionally*, you can get kicked off

What if the scientific community (reviewers) complain?

Some scientists complain that MTurk is not a legitimate data collection platform

Luckily, MTurk has been studied extensively

Overall findings: MTurk is at least as good as other recruitment tools (e.g. Qualtrics, IPSOS)

…and better than using student samples

What if the scientific community (reviewers) complain?

Works to cite:

- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057
- Boas, T. C., Christenson, D. P., & Glick, D. M. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods*, *8*(2), 232–250. https://doi.org/10.1017/psrm.2018.28
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980
- Kang, R., Brown, S., Dabbish, L., & Kiesler, S. B. (2014). Privacy Attitudes of Mechanical Turk Workers and the US Public. *SOUPS*, 37–49. https://www.usenix.org/sites/default/files/soups14_proceedings.pdf#page=44

What if the scientific community (reviewers) complain?

Works to cite (cont'd):

- Kelley, P. G. (2010). Conducting Usable Privacy & Security Studies with Amazon's Mechanical Turk. *Proceedings of the SOUPS 2010 Usable Security Experiment Reports Workshop*. http://cups.cs.cmu.edu/soups/2010/user_papers/Kelley_mTurk_USER2010.pdf
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 453–456. https://doi.org/10.1145/1357054.1357127
- Lowry, P. B., D'Arcy, J., Hammer, B., & Moody, G. D. (2016). "Cargo Cult" science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels. *The Journal of Strategic Information Systems*, *25*(3), 232–240. https://doi.org/10.1016/j.jsis.2016.06.002

What if the scientific community (reviewers) complain?

Works to cite (cont'd):

- Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.
- Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y
- Redmiles, E. M., Kross, S., & Mazurek, M. L. (2019). How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. *2019 IEEE Symposium on Security and Privacy*, 1326–1343. https://doi.org/10.1109/SP.2019.00014
- Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in mechanical turk. *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 2863–2872. https://doi.org/10.1145/1753846.1753873

What if the scientific community (reviewers) complain?

From Buhrmeister et al. (2011):

Findings indicate that:

- MTurk participants are slightly more demographically diverse than are standard Internet samples and are significantly more diverse than typical American college samples
- participation is affected by compensation rate and task length, but participants can still be recruited rapidly and inexpensively
- realistic compensation rates do not affect data quality
- the data obtained are at least as reliable as those obtained via traditional methods.

Note: payment standards have gone up since this research came out!

What if the scientific community (reviewers) complain?

From Lowry et al. (2016):

MTurk is a convenience sample:

- Also true for institutional samples
- MTurk is more diverse and can be targeted/screened

Unknown participants:

- Always true except when participation is non-anonymous*
- MTurk population is studied extensively

Repeat users may distort results:

- They can be filtered

Users can lie and cheat:

- Always a problem
- Dishonest workers get reduced HIT rate, which is bad for them

What if the scientific community (reviewers) complain?

From Lowry et al. (cont'd):

Users may not pay full attention:

- Always a problem
- Make sure to have plenty of attention checks!

People who do surveys for pay are not "normal":

- Neither are most other samples

Data quality issues:

- Mturk samples can easily be cross-validated
- Now possible to do truly anonymous studies (avoid social desirability issues)

Inability to contextualize research:

- Workers can be filtered if needed

Alternatives

to Amazon Mechanical Turk

Craigslist

- You can post a link to your study in jobs>etcetera
  - Commonly accepted place to post odd jobs
- You pay to post your ad, per city/region
  - In big cities you can get off the front page very quickly, pay again to resurface
- Payment is more difficult
  - Tip: use a raffle
  - Tip: pay extra for quick participation; pay for referrals

Prolific

- Very similar to MTurk
  - Overhead is also 40%
  - But has good participants outside the US (mostly UK and Europe)
- **When recruiting EU participants, beware of GDPR!**
  - IRB consent form mostly covers regulations
- Will discuss differences with MTurk here
  - More details: http://bit.ly/prolificSetup and https://researcher-help.prolific.co/hc/en-gb

Prolific

Connection to external survey is different

At the beginning of your study, you must save the participant's Prolific ID to your DB for identification

At the end of your study, you must give participants a "completion code" (provided by Prolific) for verification

Prolific

Saving the participant's Prolific ID:

- Option 1: get it from the URL parameters
  - Prolific sends pps to qualtrics.com/yourstudy?ID=xxxx
  - You can capture this ID automatically
- Option 2: have a question field for it early in your survey
  - Participants are used to getting this question and likely have their ID memorized

Giving participants the completion code:

- Option 1: link to Prolific using URL parameters
  - E.g. app.prolific.co/submissions/complete?cc=xxxx
- Option 2: display the code on screen
  - Participants will have to copy-paste it

Prolific

Prolific has advanced filters

You can select participants based on nationality, language skills, gender, age group, product ownership, etc.

- Use with caution: users may forget to update this data
- You may not use additional filters in your own survey
- Selecting multiple criteria uses a logical AND; selecting multiple options per criteria uses a logical OR
- Tip: if you want to e.g., a gender-balanced sample: create separate studies with a specific N for every gender option

Just like on MTurk, you can prevent repeat participating using a "custom block list"

Prolific

Payment is slightly different

MTurk only has pay or reject; Prolific allows you to contact participants

- You can ask the participant to "withdraw" their submission
- This does not affect their reputation

Questions?

Slides at http://bit.ly/RCRslides