Data quality challenges when collecting data using Amazon's Mechanical Turk

Jesse Chandler

What is Mechanical Turk?

Simple. Security, recruitment, identity verification and payment handled by Amazon
Fast. Hundreds of responses per day
Elastic. Scales up and down quickly
Cost Effective. ~$0.14 per respondent minute
Quality is generally high

- Diverse
- Decent reliability
- Results (usually) agree with studies conducted using other samples
- Widespread impact on the social sciences

Coppock, 2017

Online panels and two stage selection

- The World
- The Internet
- MTurk
- Your Study
Challenge 1: Sampling

Convenience samples are accidental samples

55% Female
N = 3010

43% Female
N = 2776
Sample composition changes over time

Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection

Logan S. Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z. Strolovitch

Time of day effects

Difallah, Filatove & Ipeirotis, 2018
Solutions

Consider how posting strategy aligns with your desired sample
Document when and how studies were posted
Lots of work to be done about what attracts people to specific studies
Be careful about generalizing findings from pilot data

Challenge 2: Experienced Participants
How many workers are there?

Ever-registered users (500k+; Amazon)

Active workers (Robinson et al., 2019)

How many workers do labs reach? (7-15k, Stewart et al. 2015)

NOTE: Higher pay decreases reach

NOTE: Larger samples increase reach

Experienced workers are very active

Robinson, Rosenzweig, Moss & Litman 2019; Chandler, Mueller & Paolacci 2013
Experienced workers are different

Demographically different
  Less educated (market research panels)
  Older (market research panels)
More financially motivated
  Self-report (MTurk and panels)
  More likely to be unemployed (MTurk and market research panels)
  Lower income (MTurk and market research panels)

Casey et al., 2017; Whitsett, 2013; Deetlefs, Chylinski & Ortmann 2015; Zhang et al., 2019

Experienced workers are experienced

Ability (CRT)

Memory (Spacing Task, Serial Position Effect; Zwann et al., 2017)

Response speed (Sequential Priming, Repetition Priming, Shape Simulation; Zwann et al., 2017; Deetlefs, Chylinski & Ortmann 2015)

Important: A mix of experienced and non-experienced workers will increase within-condition variance in between-subjects designs
Experience and vanishing effects

Rand et al., 2014

Experience and forgetting

Chandler et al., 2015
Solutions

Manage who participates in studies using whitelists and blacklists
- Those known to have participated in similar studies
- Those who have completed many studies

Use novel study materials

Note that experience isn’t always a bad thing!

Challenge 2: Fraud
Do you have a child with autism?

Demographic Question (N = 1201) Prescreening Question (N = 1196)

People respond to incentives

Fraud Rate

<table>
<thead>
<tr>
<th>Incentive</th>
<th>Fraud Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 cents/minute (N = 279)</td>
<td>5.7%</td>
</tr>
<tr>
<td>20 cents/minute (N = 549)</td>
<td>15.8%</td>
</tr>
</tbody>
</table>

Chandler & Paolacci, 2017
Fraud causes spurious individual differences

Framing Sharpe, Huber & Netzer, 2017

Experience revisited – do you own a VR headset?

First data collection (N = 365) One week later (N = 203) Five weeks later (N = 162)

Chandler & Paolacci, 2017
Solutions

Technical solutions are imperfect (but help)

Do not explicitly list prescreening requirements

Do not screen out ineligibles: route eligible to a second survey and pay them a bonus for completing

Record worker IDs in surveys so you can link survey attempts together

Challenge 3: “Bots” and satisficing
Maybe “survey farmers”

Eggplant or Brinjal?

![Eggplant image]

Litman: https://tinyurl.com/y69k37x8

VPS and Data Quality Problems

Sample $N = 2354$

- Sleep $>24$ hrs at a time or only every 3 days
- Read $>10$ wps
- Suspicious IP addresses

$n = 114$  
$n = 36$  
$n = 198$  
$n = 234$

$n = 177$  
$n = 252$  
$n = 629$  
$n = 377$

$n = 75$  
$n = 90$
So what?

Introduces noise
Lowers reliability
Reduces correlations
Increases between group variance

Impact on correlations

Chandler, Sisso & Shapiro, in press
Measuring rare events

- Hallucinations and Violent Behavior: $r = 0.82$
- Autism Spectrum and Violent Behavior: $r = 0.20$
- Depression and Vaccine Recency: $r = 0.51$

- Hallucinations and Violence: $r = 0.16$
- Autism Spectrum and Violence: $r = -0.09$
- Depression and Vaccine Recency: $r = 0.27$

Chandler, Sisso & Shapiro, in press

Making a market for lemons

Most workers have a 99% approval rate (Matherly, 2019)

A substantial proportion of workers originating from suspicious IP addresses have good reputations:

<table>
<thead>
<tr>
<th></th>
<th>95-98%</th>
<th>99%+</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-1000</td>
<td>41%</td>
<td>18%</td>
</tr>
<tr>
<td>1000+</td>
<td>49%</td>
<td>14%</td>
</tr>
</tbody>
</table>

Sisso unpublished data
Data filtering

Chandler, Sisso & Shapiro, in press

Panel management

Maintain blacklists of low quality workers
Reject payment where appropriate
Block workers where appropriate
Not clear what “appropriate” means
Closing thoughts

MTurk is low cost for a reason

Schwarz (2007): “The initiators of the first Cognitive Aspects of Survey Methodology conferences hoped to build a two-way bridge between psychology and survey methods to facilitate an exchange that would advance basic research and improve survey practice. In the two decades since these conferences, this bridge has seen considerable traffic. However, much of this traffic has been from psychology to survey methods.”

Some hope still that traffic can include substantive contributions to psychological theory

Need for “another lane” on the bridge to transfer knowledge about how to improve methodological rigor

Questions?
Contact: JChandler@Mathematica-mpr.com
Selected References

An overview of MTurk


Sampling


Fraud


"Bots" (but – probably – not Bots)
