Automated, web-based data collection using IbexFarm and Amazon’s Mechanical Turk

Joshua Levy
9/25/13
Typical Data Collection

• IRB approval is sought and granted
• Subject and experimenter rendezvous at testing location
• Experimenter administers informed consent procedures
• Task instructions presented
• Experimenter supervises data collected
• (Experimenter briefs subject on nature of study)
• Subject is compensated
Typical Data Collection vs. Web-based Data Collection

• IRB approval is sought and granted
• Subject and experimenter rendezvous at testing location
• Experimenter administers informed consent procedures
• Task instructions presented
• Experimenter supervises data collected
• (Experimenter briefs subject on nature of study)
• Subject is compensated
Web-based Data Collection

- IRB approval is sought and granted
- Subject is stationed at computer with internet access
- Subject indicates informed consent in absence of experimenter (minimal risk studies)
- Task instructions presented
- Automated & unsupervised data collection
- (Automated briefing of subject on nature of study)
- Subject is compensated
Types of Data that can be collected via the web

- Judgment (Likert or otherwise)
- Forced-choice response
- Self-paced reading
- Web-based eye-tracking
  - Low spatial accuracy
- RT
  - RTs sent to server are unreliable
Types of Stimuli that can be presented via the web

• Text
  – **IbexFarm** (by Alex Drummond)

• Auditory and/or Images
  – Directly through Mechanical Turk or through other software
MTurk

• Amazon’s Mechanical Turk is portal where workers (subjects) find experiments posted by requesters (experimenters)
• MTurk is also secure portal for sending money to subjects as compensation
• Nomenclature:
  – HIT = job / set of items that subjects see
  – Assignments = number of iterations for given HIT
• Market Rate: $0.02 per single judgment (Sprouse, 2011)
MTurk

- MTurk allows specification of custom worker requirements by many factors, including:
  - Participation rate
  - Approval rate
  - IP address Location

![MTurk interface](image-url)
Turker Demographics (Mason & Suri, 2012)

• More than 100,000 workers from 100 countries (*NY Times*, 2007)
  – Majority of workers from US (46.8%) and India (34%) (*Ipeirotis*, 2010)
  – Payment in USD and INR only

• 55% Female; 45% male (87.5% of Turkers)
Turker Demographics
(Mason & Suri, 2012)

Mean = 32
Median = 30
Turker Demographics (Mason & Suri, 2012)

Compiled income data from several MTurk studies (each binned income in different ranges)

Fig. 2 Distribution of the maximum of the income (in U.S. dollars) interval self-reported by workers
How do Turkers Compare to Traditional Subjects?

• Varied population

• Qualitatively similar behavior

• Qualitatively different behavior
  – Category learning (Crump et al., 2013)
How do Turkers Compare to Traditional Subjects?

Category Learning
Shepard et al. (1961)
How do Turkers Compare to Traditional Subjects?

Figure 8. Cognitive Learning: A comparison between the learning curves reported in Nosofsky et al. (1994) data and the AMT replication data in Experiment 8. The probability of classification error as a function of training block. The top panel shows the learning curves estimated by Nosofsky et al. [38] using 120 participants (40 per learning problem) who each performed two randomly selected problems. The right panel shows our AMT data with 228 participants, each who performed only one problem (38 per condition). We ended the experiment after 15 blocks, although Nosofsky et al. stopped after 25. Thus, the Nosofsky et al. data have been truncated to facilitate visual comparison.
How do Turkers Compare to Traditional Subjects?

Figure 9. Cognitive Learning: The average number of block to criterion for each problem, an index of problem difficulty. The average number of blocks it took participants to reach criterion (2 blocks of 16 trials in a row with no mistakes) in each problem. The white bars show the estimated average number of blocks to criterion reported by Nosofsky et al. [38].
How do Turkers Compare to Traditional Subjects?

• Sprouse (2011) presented subjects with sentences and asked subjects to rate sentences using magnitude estimation, relative to a standard

• Standard and modulus repeated every seven trials (always visible on screen)

Standard: Who said my brother was kept tabs on by the FBI? 100

Item: What did Lisa meet the man that bought? ___
How do Turkers Compare to Traditional Subjects?

• Sprouse (2011) presented subjects with sentences and asked subjects to rate sentences using magnitude estimation, relative to a standard

• 176 subjects in traditional data collection course credit or $5
• 176 subjects in Mturk data collection $3
How do Turkers Compare to Traditional Subjects?

• Sprouse (2011) presented subjects with sentences and asked subjects to rate sentences using magnitude estimation, relative to a standard

• 96 items presented in one of four lists
• Items were island violations or non-violating controls
• Center embedding, comparative, & agreement attraction illusions
How do Turkers Compare to Traditional Subjects?

Standard:

Who said my brother was kept tabs on by the FBI?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(2)</td>
<td><em>Whether Island Effect</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What do you think that John bought?</td>
<td>(control)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*What do you wonder whether John bought?</td>
<td>(violation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td><em>Complex Noun Phrase Island Effect</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What did you claim that John bought?</td>
<td>(control)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*What did you make the claim that John bought?</td>
<td>(violation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td><em>Subject Island Effect</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What do you think interrupted the TV show?</td>
<td>(control)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*What do you think the speech about interrupted the TV show?</td>
<td>(violation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td><em>Adjunct Island Effect</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What do you think that John forgot at the office?</td>
<td>(control)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*What do you worry if John forgets at the office?</td>
<td>(violation)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How do Turkers Compare to Traditional Subjects?

Standard:
Who said my brother was kept tabs on by the FBI?

(6) Center Embedding Illusion

*The ancient manuscript that the grad student who the new card catalog had confused a great deal was studying in the library was missing a page. (violation)
?The ancient manuscript that the grad student who the new card catalog had confused a great deal was missing a page. (illusion)

(7) Comparative Illusion

*More people have graduated law school than I have. (violation)
?More people have been to Russia than I have. (illusion)

(8) Agreement Attraction Illusion

*The slogan on the poster unsurprisingly were designed to get attention. (violation)
?The slogan on the posters unsurprisingly were designed to get attention. (illusion)
How do Turkers Compare to Traditional Subjects?
How do Turkers Compare to Traditional Subjects?
How do Turkers Compare to Traditional Subjects?

• Sprouse (2011) found MTurk outlier rate (14.2%) higher than traditional outlier rate (1.7%)

• Also, higher incidence of subject attrition due to participation in HITs that subjects freely admit they are not eligible for

• Recommendation: Run more subjects and include a questionnaire about relevant inclusion criteria
Conclusions

• Automated data collection is time efficient
• Allows for broader sample
  – Age
  – Geography
  – Language
  – Income?
• Validity of web-based data compared to traditional data collection is mixed, but positive for acceptability judgment tasks