Determining Amazon's Mechanical Turk's Survey Response Consistency: A UROP Project

Jerika Eppel, Serguei Pakhomov, Raymond Finzel, Michael Kotlyar

Introduction

Amazon's Mechanical Turk (AMT) could provide researcher's a low-cost way of crowdsourcing reliable and diverse data in an accessible, low-cost format. The platform has many benefits for traditional research, including:

- Larger and more diverse demographics than college campuses, which reduces participant bias.
- Lower cost than traditional methods of data collection.
- Faster experimental iterations which allow studies to move through pilots faster.

Prior research has identified several concerns with AMT as a platform.

- Compensation being tied to surveys-per-hour could incentive AMT participants to complete surveys inaccurately to finish them faster.
- Prior research in attentiveness and data quality has been positive, however, several studies did measures with traditional attentiveness markers. Due to the volume of surveys a participant completes through AMT, they could learn how to identify these markers and adjust their responses.
- Studies often limit AMT participants to those with a high "reputation" score, which results in minimal research on low "reputation" score participants. This limitation can extend the time needed to recruit participants.

Rational

Based on the premise that consistency in responses to survey questions reflect the overall reliability of an AMT participant's survey results, we hypothesized that a high proportion of survey responses regarding lifestyle and smoking behaviors would not be reliable.

Methods

- AMT participants that identified as from the United States, over 18, and, after 100 participants, as a smoker were asked to complete a health behaviors survey for a longitudinal research study. Participants were paid $0.50 for their time.
- The survey was conducted through RedCap and had a maximum of 65 questions mapped across demographics, health overview, physical activity, smoking behaviors, e-cigarette perceptions, and nutrition. The number of questions varied with conditional logic embedded in the survey.

Answer sets were mapped afterward to track the following three types of inconsistencies:

- Dependent Response Redundancy: two or more questions that have overlapping answers.
  - i.e. "Have you smoked in the last 30 days?" and "When is the last time you had a cigarette?"
- Single response inconsistency: question's response is impossible.
  - i.e. "How much did you pay for your last pack of cigarettes?"
- Patterned response: Responses follow a pattern across fundamentally different questions.
  - i.e. Selecting the first radio button in each group on an entire page.

Several sets of questions had more than 1 answer set that could lead to inconsistency and those were counted separately for analysis.
Traditional reliability measures such as attention checks were not used in this survey.
The average time of surveys was measured from the completion of the consent to the completion of the final survey question.

Results

- 541 AMT participants consented to screening survey and 398 completed the entire survey.
- Participants took between 1:18 and 49:26 minutes to complete the survey with an average time of 6.99 minutes.
- 18 answer sets were identified as inconsistent with 7 being dependent response redundancy, 2 being single response inconsistency, and 7 being pattern responses.
- Each inconsistency flagged between 2 and 9 subjects with one outlier flagging 36. This outlier was removed after being determined as a survey design miscommunication.
- Without the outlier, 35 surveys had 1 inconsistency, 4 surveys had 2 inconsistencies, and no surveys had over 2 inconsistencies. The maximum number of inconsistencies possible was 11.

Discussion

- A majority of respondents were consistent within the survey, even when it wasn't restricted to high-reputation workers only.
- There was only a minimal time difference (20 seconds) between consistent and inconsistent surveys, meaning participants weren't motivated to be inconsistent in order to complete the survey faster.
- AMT participants consistency did not seem to be affected by the lack of attention checks or other traditional consistency measures, signally that they aren't "learning" how to spot these measures and adapting their responses around them.

Limitations

- $0.50 compensation might have influence participant's accuracy. Previous studies that had been studied had much lower compensation ($1.66 an hour), whereas this survey was an average of $4.30 an hour.
- Participants might have misunderstood the survey questions instead of purposefully responding with inconsistency.
- The number of inconsistent surveys were much lower than consistent surveys, thus the demographics of inconsistent surveys might not be as accurate of the sample.
- Since traditional reliability measures were not used, there is little research into the reliability of the consistency measures that were used.

Response Breakdown

Inconsistency	Number of flagged participants
Dependent redundancy	
Smoking history and recent smoking behaviors 1	8
answer set 2	1
Smoking frequency in the last month	3
Disjointed free write response	7
Today's age verses Smoking age	1
Smoking age – smoking frequently and daily	4
Quit smoking frequency	3
Single inconsistency	
Cigarette price	5
Time sitting per day	34 [Outlier]
Pattern responses	
Physical Activity – Days per week 1	2
answer set 2	1
answer set 3	1
answer set 4	1
Physical Activity – Time per day	3
Nutrition – Per day – answer set 1	1
answer set 2	2

Demographics Comparison

Metric	Consistent-only surveys	Inconsistent-only surveys
Population	511	35
Completers	364 completers; 147 non-completers	34 completers; 1 non-completer
Average Age	35.3	33.7
Sex	51.5% male; 48.5% female	54.3% male; 45.7% female
Native Language	92.9% English; 7.1% other	94.3% English; 5.7% other
Highest level of education	1% Some high school; 12.7% High school diploma or equivariant; 37.8% Some College; 34% Undergraduate College Degree; 14.5% Graduate/Professional Degree	2.9% Some high school; 14.3% High school diploma or equivariant; 37.1% Some College; 37.1% Undergraduate College Degree; 8.6% Graduate/Professional Degree
Completed surveys average time	7.04 minutes	6.84 minutes