Screening Participants on Mechanical Turk: Techniques and Justifications

Julie S. Downs
Mandy Holbrook
Emily Peel
Carnegie Mellon University
Background

• mTurk data: perception of poor quality
  – Concern about “non-participants”
• Multiple strategies to screen them out
  – Drop fast performing participants
  – Force slower speeds
  – Gold standard questions
  – Instructional manipulation check
• Not much consensus
Goal

• Assess strategies for restricting data collection and data retention
• Suggest the most beneficial techniques to:
  – Identify noisy data
  – Omit the fewest responses possible
  – Prevent creating a biased sample
Design and Procedure

• Five surveys (N= 1,888)
  – Demographics: 40% male, 37 years old (SD=11.98, Range= 61)
  – Different selection of screening and criterion variables
  – Posted at different times on AMT
• Participants were only located in the United States
• 95% approval rating
• Required 500+ HITs*
• Pay estimated $8/hour
• Bonus in some cases

*Except where noted
Screening Techniques Tested

• Meta-data from typical surveys
  – Time on task
  – Length of open-ended feedback

• Responses that identified “poor participants”
  – Gold standard questions
    • Objectively correct answer
  – Instructional manipulation check
  – General knowledge questions
    • Pre-selected to be extremely easy (Moore, 2010)
Criterion Variables

• Reliability measures
  – Classic individual difference scales
    • Social Desirability, Need for Cognition, Self Esteem (Cronbach’s α)
  – Internal Consistency (percentage)
    • Violations of dominance axiom

• Effect sizes for classic biases
  – Cognitive performance and biases in judgment and decision making
    • Stroop Task, Framing Effect, Sunk Cost, Risk Aversion Bias, Uncertainty Bias
Reliability Measure Tasks

• Measure 1: Self Esteem Scale
  – Measures an individuals self esteem at the moment

On the following few pages there is a list of statements dealing with your general feelings about yourself. Please indicate the extent to which you agree or disagree with the statements.

All in all, I am inclined to feel that I am a failure.

I certainly feel useless at times.

I wish I could have more respect for myself.
Reliability Measure Tasks

• Measure 2: Social Desirability Scale
  – Assesses whether respondents are responding truthfully or are misrepresenting themselves in order to manage their self-presentation
Reliability Measures Task

• Measure 3: Need for Cognition Scale
  – Measures the tendency for an individual to engage in and enjoy thinking

<table>
<thead>
<tr>
<th>Describe the extent to which you agree with the following statements.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would prefer complex to simple problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I like to have the responsibility of handling a situation that requires a lot of thinking.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thinking is not my idea of fun.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reliability Measures Task

- **Measure 4: Internal Consistency**
  - Every gamble choice was a choice: risk of $100 and a sure amount
  - When offered a gamble or a sure win of $50, and the subject chose the sure win, then subjects should definitely chose sure win $60
  - If they don’t, they have violated dominance
Classic Bias Tasks

- Bias 1: Stroop Task
  - Two different tasks (counterbalanced)
    - Asked to select the color of the letters “XXXXXX” were printed in
    - Asked to select the color the words were printed in
  - Naming the color of the word takes longer and more prone to errors
Classic Bias Tasks

• Bias 2: Framing Effect
  – The two options have the same `objective' product of probabilities and outcomes
    • When framed as deaths- prefer the risky option
    • When framed as saving- prefer the safer option
Classic Bias Tasks

- Bias 3: Sunk Cost
  - 1 of 2 scenarios are presented, only 1 involves a sunk cost
  - Respondents who received sunk cost question are likely to continue construction because millions have been invested, a psychological sunk cost

**Scenario 1:**
As the president of an airline company, you have invested 10 million dollars of the company's money into a research project. The purpose was to build a plane that would not be detected by conventional radar, in other words, a radar-blank plane. When the project is 90% completed, another firm begins marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company is building. The question is: should you invest the last 10% of the research funds to finish your radar-blank plane?

**Scenario 2:**
As president of an airline company, you have received a suggestion from one of your employees. The suggestion is to use the last 1 million dollars of your research funds to develop a plane that would not be detected by conventional radar, in other words a radar-blank plane. However, another firm has just begun marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company could build. The question is: should you invest the last million dollars of your research funds to build the radar-blank plane proposed by your employee?
Classic Bias Tasks

- Bias 4: Risk Aversion Bias
  - To calculate, the number of gambles taken on wins and losses were totaled separately
  - Those who took more gambles on wins than losses were risk averse
Classic Bias Tasks

- Bias 5: Uncertainty Bias
  - When precursor is unknown people have a harder time making later decisions
  - Used “doesn’t matter” as a proxy that it was hard to make this decision
  - The bias would predict choosing doesn’t matter more often when first flip is unknown

In each of the following questions, a coin will be flipped to see if you get a choice or not. Without knowing the result of the first flip, what would you choose in each of the following situations?

First Flip:

Flip First Coin
  - If Heads, get the Choice below
  - If Tails, don’t get the Choice below, win $0

Choice: Before the first flip, which of the following do you like best:

- Flip Second Coin
  - If Heads, win $100
  - Otherwise, win $0

- Sure Win
  - Win $50 for sure

- Doesn’t Matter to Me
Ideal Screening Tool Properties

**Scale Reliabilities**

- **Self-Esteem**: Cronbach's Alpha = 0.93
- **Social Desirability**: Cronbach's Alpha = 0.84
- **Nd Cognition**: Cronbach's Alpha = 0.95
- **Consistency**: Cronbach's Alpha = 0.93

**Bias Effect Sizes**

- **Stroop**: Effect Size = 0.15
- **Sunk Cost**: Effect Size = 0.51
- **Loss Aversion**: Effect Size = 0.58
- **Uncertainty**: Effect Size = 0.78
- **Framing**: Effect Size = 0.09

* if n<5, no analyses were performed
Instructional Manipulation Check

Sports Participation

Most modern theories of decision making recognize the fact that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables, can greatly affect the decision process. In order to facilitate our research on decision making, we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. So, in order to demonstrate that you have read these instructions, please ignore the sports items below. Instead, simply select the title at the top of this screen (i.e., "sports participation"), and then click to the next screen. Thank you very much.

Which of these activities do you engage in regularly? (check all that apply)

- Skiiing
- Soccer
- Snowboarding
- Running
- Hockey
- Football
- Tennis
- Basketball
- Cycling
Instructional Manipulation Check

Scale Reliabilities

<table>
<thead>
<tr>
<th></th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Esteem</td>
<td>Pass: 0.93</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.92</td>
</tr>
<tr>
<td>Social Desirability</td>
<td>Pass: 0.94</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.89</td>
</tr>
<tr>
<td>Consistency</td>
<td>Pass: 0.84</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.75</td>
</tr>
</tbody>
</table>

Bias Effect Sizes

<table>
<thead>
<tr>
<th></th>
<th>Effect Size of Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroop</td>
<td>Pass: 0.13</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.18</td>
</tr>
<tr>
<td>Sunk Cost</td>
<td>Pass: 0.22</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.53</td>
</tr>
<tr>
<td>Loss Aversion</td>
<td>Pass: 0.16</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.61</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>Pass: 0.08</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.78</td>
</tr>
<tr>
<td>Framing</td>
<td>Pass: 0.02</td>
</tr>
<tr>
<td></td>
<td>Fail: 0.74</td>
</tr>
</tbody>
</table>
Timing to Finish Survey

Time to Complete Survey #4

Mean = 712.65
Std. Dev. = 416.373
N = 312

Frequency

Seconds
Finished Survey Too Quickly
Open-Ended Responses

What did you like most about this survey?

What did you like least about this survey?

Is there anything you would recommend to make it better?
Short Open-Ended Responses

Scale Reliabilities

- Self-Esteem
  - Included: 0.93
  - Omitted: 0.94
- Social Desirability
  - Included: 0.84
  - Omitted: 0.84
- Nd Cognition
  - Included: 0.96
  - Omitted: 0.92
- Consistency
  - Included: 0.94
  - Omitted: 0.90

Bias Effect Sizes

- Stroop
  - Included: 0.15
  - Omitted: 0.23
- Sunk Cost
  - Included: 0.22
  - Omitted: 0.53
- Loss Aversion
  - Included: 0.10
  - Omitted: 0.68
- Uncertainty
  - Included: 0.78
  - Omitted: 0.87
- Framing
  - Included: 0.10
  - Omitted: 0.13
Extremely Easy Questions

Please answer the following 5 general knowledge questions.

An octopus has how many arms?

Cinderella, The Little Mermaid, Aladdin, and The Lion King are all films produced by what famous entertainment company?

Singer Celine Dion sang the hit song "My heart will go on" for the soundtrack of what 1997 film about the sinking of a famous ship?

What is the capital of the United States?

What is the major pumping organ of the human circulatory system?
Missed Extremely Easy Questions

Scale Reliabilities

- Self-Esteem: Included (n=447) - 0.93, Omitted (n=25) - 0.93
- Consistency: Included (n=447) - 0.93, Omitted (n=25) - 0.94

Bias Effect Sizes

- Stroop: Included (n=447) - 0.12, Omitted (n=25) - 0.12
- Sunk Cost: Included (n=447) - 0.52, Omitted (n=25) - 0.52
- Loss Aversion: Included (n=447) - 0.60, Omitted (n=25) - 0.61
- Uncertainty: Included (n=447) - 0.79, Omitted (n=25) - 0.78

Cronbach’s Alpha

Effect Size of Bias
Thank you!

You will need to enter **two codes** indicating that you have completed the survey.

The first is a unique number that will let us check your data to make sure that you completed the full survey:
Here is your first code: **55398**

For the second code, despite Amazon’s attempts to ensure the intelligence tasks are done by humans, we are concerned with the possibility that computers may be taking this task, which would add significant noise to our research data.

If you are a computer enter the code: **184ckdD**
If you are a real person, enter the word that is the opposite of "white" followed by the opposite of "dog"
Missed Gold Standard Questions

Scale Reliabilities

<table>
<thead>
<tr>
<th>Scale</th>
<th>Included (n=396)</th>
<th>Missed 1 (n=86)</th>
<th>Missed 2 or more (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Esteem</td>
<td>0.94</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>Consistency</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bias Effect Sizes

<table>
<thead>
<tr>
<th>Bias Effect</th>
<th>Included (n=396)</th>
<th>Missed 1 (n=86)</th>
<th>Missed 2 or more (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroop</td>
<td>0.11</td>
<td>0.20</td>
<td>0.35</td>
</tr>
<tr>
<td>Sunk Cost</td>
<td></td>
<td>0.39</td>
<td>0.49</td>
</tr>
<tr>
<td>Loss Aversion</td>
<td>0.32</td>
<td>0.48</td>
<td>0.62</td>
</tr>
<tr>
<td>Uncertainty</td>
<td></td>
<td>0.59</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Conclusions

• Reliabilities remain high
  – Most were well above acceptance threshold
  – Lowest was 0.75

• Mixed results for biases
  – Some increase and some decrease
  – Screening techniques may change sample

• Is screening justified?
  – Undermines benefit of mTurk: public, diversity

• Danger of undisclosed flexibility in analysis
  – Especially if screening decision are made post hoc