Amazon Mechanical Turk Workers Can Provide Consistent and Economically Meaningful Data

David Blake Johnson
University of Central Missouri

John Barry Ryan
Stony Brook University

Motivation

Amazon Mechanical Turk (AMT) has become a common tool for researchers in social sciences because researchers are looking for samples that are:

- More heterogeneous than student samples and,
- Inexpensive.

Critics of AMT worry that AMT samples are unreliable and of poor quality.

- Formal criticisms seen academic literature:
  - Frequency of participation
  - Unrepresentative samples
- Informal criticisms seen in ref. reports and editorial letters:
  - Loss of experimental control
  - Low incentives
  - Overall subject quality

Given that AMT experiments generally replicate the results using nationally representative and laboratory studies (c.f., Horton et al., 2011; Amir et al., 2012; Mullinix et al., 2015; Gibson and Johnson, 2018; Arechar et al., 2018), these criticisms seem puzzling.

Methods and Data

We explore the consistency of the characteristics reported by AMT workers and the relevancy of the characteristics. Data comes from 11 different experiments run on AMT over a 6 year period. Data sets generated using AMT include workers' identification numbers (workerid) and the date the experiment was posted.

- Allows us to compare respondents' first responses to future responses.

Experiments are very different but all ask subjects to report their Age, Gender, Impulsivity, and Subjective Risk Preferences

- Impulsivity measured using a multi-item index (Barrett Impuslivity Scale)
- Subjective risk preferences measured using the English translation of the German SOEP question (popularized in Dohmen et al., 2011).
  - *How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?*
    - Larger values indicating a greater willingness to take risk.

Use data from experiments that ask workers to report at least two of the measures above. 5,347 subjects completed at least one experiment. 3,566, 730, and 223 completed 1, 2, and 3 or more experiments, respectably.

Age and Gender Consistency

Workers who complete more than one experiment are more likely to report being male, less impulsive, and less willing to take risk.

- 666 of 778 workers consistently report their gender (98.23%)
- 418 workers report their age more than once. The average reported age when workers completed their first(last) experiment 33.12(33.97) which translates into a difference of .85 years.
  - Difference is not significantly different from .793 which is the average time between experiments (% year).

Age and Gender are easy to consistently answer so these variables are a "low-bar" test.

- Perform more comprehensive analyses of two respondent personality traits: impulsiveness and willingness to take risk.

Risk and Impulsivity Consistency

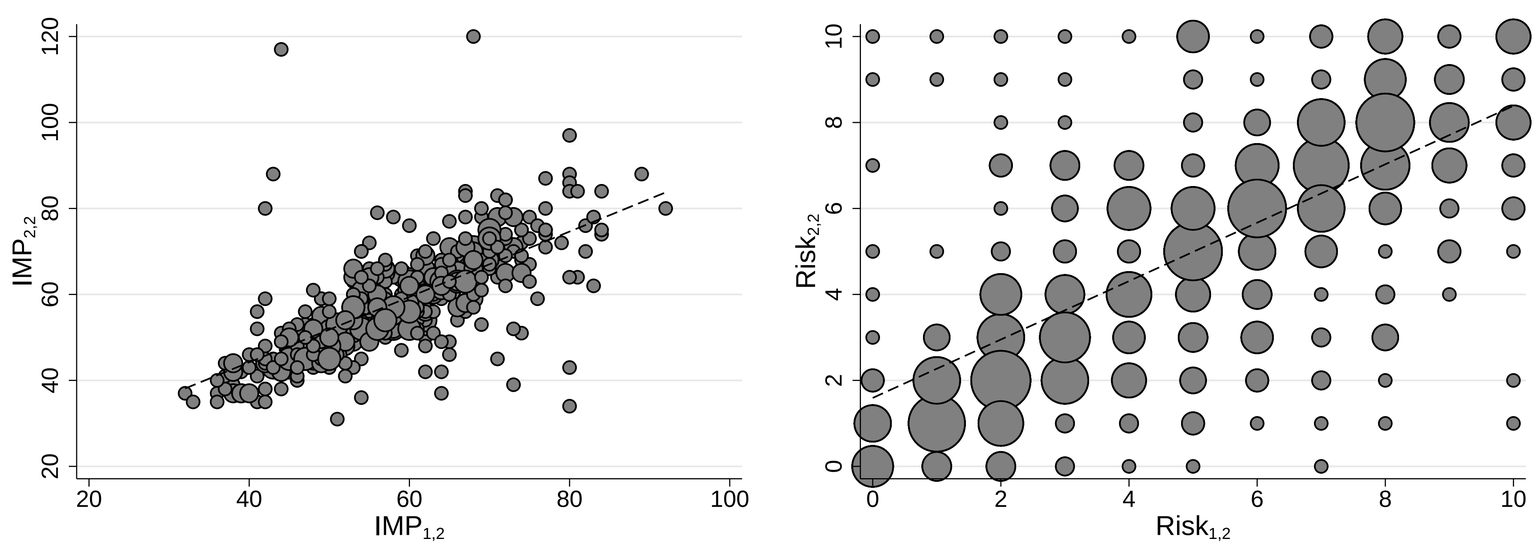
The average time difference between workers' first (58.628) and last (58.369) Impulsivity Test is 339 days (n =406). Difference in workers' first and last Impulsivity score is not statistically different ($p = 0.5843$).

- Both scores are highly correlated ($r = 0.679$, $p < 0.001$). Not as high as Stanford, et. al. (2009), reporting a *one month* correlation across test scores of 0.83, but if the 8 most inconsistent workers are removed, correlation rises to 0.804.

The average time difference between workers' first and last response to the subjective risk question is 418 days (n =553). The difference in workers' first (4.911) and last (4.929) response to the subjective risk question is not statistically different ($p = 0.852$).

- Two scores are highly correlated ($r = 0.672$, $p < 0.001$). Test and re-test correlation is higher than the 30-49 day reliability ($r = 0.60$) reported in the SOEP manual.

Figure 1: Scatter Plot of Workers First and Last Reported Impulsivity/Risk



Scatter plot of first recorded impulsivity/subjective risk ($IMP_{1,2}$/$RISK_{1,2}$) and final recoreded impulsivity ($IMP_{2,2}$/$RISK_{2,2}$).

Economic Meaningfulness of Data

Now show that subjective risk preferences correlate with decisions in a real stakes task (Johnson and Webb, 2017; Gibson and Johnson, 2018). Subjects pick 1 of the 20 lotteries (below) which is played and are paid based off of the outcome of the lottery. Riskier (safer) lottery choices are lower (higher) in index number.

Table 1: Lotteries In Johnson and Webb (2017) and Gibson and Johnson (2018)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Prob	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
Prize	5.00	4.75	4.50	4.25	4.00	3.75	3.50	3.25	3.00	2.75	2.50	2.25	2.00	1.75	1.50	1.25	1.00	0.75	0.50	0.25

In Johnson and Webb (2017) and Gibson and Johnson (2018) workers report subjective risk preferences and complete the lottery task. Models 1 and 2 (n =122) use the measures collected at the same point in time. Models 3 and 4 use the 34 respondents who participated in more than one study and uses the subjective risk gathered in the first study the subject participated in (i.e., before the lottery experiment).

Table 2: Risk Preferences and Risky Decision Making in a Real Stakes Task

	Model 1		Model 2		Model 3		Model 4	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Contemporaneous Risk	**-0.579**	0.151	**-0.562**	0.159				
1st Risk Response					**-0.733**	0.229	-0.800	0.619
Days Since Response					0.003	0.003	0.002	0.006
Days X 1st Response							0.000	0.001
Constant	14.404	0.835	12.319	2.817	14.333	2.294	14.678	3.751
Controls			✓				✓	

Find an inverse correlation between workers' responses to the subjective risk question and the lottery selected. This suggests that the responses to the subjective risk question are consistent and economically meaningful. Model 3 shows the relationship holds – even if the subjective response was from an earlier time point. Model 4 shows the magnitude of the relationship does not differ as the number of days since the first subjective response increases.

Conclusion

- Demonstrate that even in inconsistent settings, with low stakes, in an uncontrolled environment, with few qualifications necessary for inclusion, workers on AMT can provide consistent and economically meaningful data.
- This suggests that the discipline should not dismiss studies using online samples in all samples as the quality of experimental data online samples produce is higher than many seem to believe.

Contact Info

- David B. Johnson
  - djohnson@ucmo.edu
- John Ryan
  - john.ryan@stonybrook.edu

LaTeX TikZposter