# Amazon Mechanical Turk Workers Can Provide Consistent and Economically Meaningful Data

November 12, 2018

**Abstract**

We explore the consistency of the characteristics of individuals who participate in studies posted on Amazon Mechanical Turk (AMT). The primary individuals analyzed in this study are subjects who participated in at least two of eleven experiments run on AMT between 2012 and 2018. We demonstrate subjects consistently report a series of demographic and personality characteristics. Further, subjective willingness to take risk is found to be significantly correlated with decisions made in a simple lottery experiment with real stakes. This suggests the quality of data obtained via AMT is unharmed by the lack of control and low stakes.

**Key Words**: Online Experiment; Amazon Mechanical Turk; Consistency; Risk; Impulsivity
**JEL Code**: C81;C83;C99

## 1 Introduction

Amazon Mechanical Turk (also known as AMT or MTurk) has become an increasingly common tool for researchers in social sciences (e.g., Clifford, 2014; Verkoeijen and Bouwmeester, 2014; Milita et al., 2017). This is partly because researchers are looking for experimental samples that are more heterogeneous than student samples and cheaper than samples recruited by survey firms and laboratory samples. However, critics of AMT worry that AMT samples are unreliable and of poor quality. The reasoning behind such worries are often based off of some mixture of formal criticisms (e.g., frequency of participation and unrepresentative samples) that are discussed in the academic literature (c.f., Krupnikov and Levine, 2014; Huff and Tingley, 2015), and informal criticisms that crop up in blog posts and referee reports (see Searles and Ryan, 2015). Given that experiments performed on AMT 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.

In this paper, we add to the growing body of literature reporting the efficacy of AMT by demonstrating that AMT worker responses are consistent across time (and where possible at level comparable to more controlled studies) and correlated, in the expected direction, with decisions that are made when there are real stakes present. Our data is from eleven different experiments run on AMT from September of 2012 to January of 2018. In each of these experiments, we collected at least two of the following measures: age, gender, impulsivity, and subjective willingness to take risk. We take advantage of the fact that AMT data sets report workers' "worker id" numbers and use this information to track worker responses over time. We show differences in responses are small – which would be consistent with measurement error – both to single questions and a multi-item index.[1] Further,

[1]In fact, the responses to the question about risk preference are more reliable than responses to one of the long-running panel survey.

we demonstrate that self-reported risk measures are highly correlated with a financially incentivized measure of risk. This is important because it suggests the consistency is economically meaningful and not the result of some decision rule used to complete surveys more quickly – for example, always picking the first option. Because we can predict the respondent's incentivized behavior with self-reported measures taken at the same time and months or, in some cases, even years before, this suggests AMT respondents provide valid measures of variables that are important to social scientists, despite the low stakes and lack of control.

## 2 AMT and Inconsistent Responses

Before proceeding further, we would like to give a brief refresher of the labor market we are using. AMT is an online labor market made up of workers (respondents/subjects) and requestors (researchers). Requestors on AMT post human intelligence tasks (HITs) that are then completed by workers in exchange for payment. In the social sciences, these HITs are often experiments (e.g., Clifford and Gaskins, 2016; Del Ponte et al., 2017; Milita et al., 2017) and are seen as an alternative to laboratory experiments that use student samples (Goodman et al., 2013). Each HIT on AMT pays workers a fixed participation fee for completing a HIT. These fixed participation fees are typically quite small and range from \$0.05 to \$1.00. This participation fee is analogous to the show-up payments paid to subjects in laboratory economics experiments. Requestors also have the option of paying workers a bonus. A convenient feature of these bonuses is that they are individually assigned after the requestor has observed workers' responses. This means requestors can observe workers' behaviors/decisions and pay them based off of their observed behavior/decisions. Like the participation fees, these bonuses are quite small relative to laboratory payments.

Demographically, AMT workers are significantly more varied than university subject pools (Ipeirotis, 2010; Buhrmester et al., 2011; Behrend et al., 2011). For example, it is

common to observe workers who report to be as young as 18 and as old as 65. However, American workers are still younger and more ideologically liberal than the overall U.S. population (Berinsky et al., 2012). More importantly, for the purposes of this paper and others using AMT, American workers on AMT generally have a lower reported income (c.f., Paolacci et al., 2010; Berinsky et al., 2012; Levay et al., 2016) are more likely to report to be unemployed (e.g., Ipeirotis, 2010; Goodman et al., 2016).

This suggests that the money earned from completing HITs might be especially valuable to AMT workers. Since the HITs generally pay small amounts, the only way to earn even minimum wage is to complete many HITs. However, this could lead to poor responses as respondents engage in satisficing or some quick decision rule (for example, always choose the middle category). It could also lead to response instability, one of the most common issues in social science surveys (Converse, n.d.). Zaller (1992) notes that one way to cure response instability is to encourage respondents to "stop and think" prior to answering. For these AMT respondents, stopping to think is costly as time spent thinking is time not spent earning money for completing another HIT.

At the same time, survey experiments conducted on the platform have been shown to replicate the results from surveys on representative samples (Berinsky et al., 2012; Mullinix et al., 2015). For example, Horton et al. (2011) replicates three experiments: i) a prisoner's dilemma, ii) a priming experiment, and iii) a framing experiment. In each of these experiments, Horton et al. (2011) finds that AMT workers' behavior is not significantly different from laboratory counterparts.[2] This result is not unique. Amir et al. (2012) explores behavior in online dictator, ultimatum, trust, and public goods games and concludes that results "are generally consistent with what is observed in the physical laboratory" despite the low stakes and lack of control. Gibson and Johnson (2018) show that AMT workers

[2]Horton et al. (2011) also demonstrates AMT workers have upward sloping supply curves.

make decisions consistent with risk averse preferences (albeit more so than laboratory subjects) across four different incentivized lottery experiments and, just as in Dohmen et al. (2011), that their decisions made in the experiments correlate, in the expected direction, with their reported subjective willingness to take risk. Finally, Arechar et al. (2018) find that contributions made in a repeated public goods game taking place online is similar to those seen in the lab. Given that online experiments are much less expensive than their laboratory counterparts and that online results generally replicate results observed in the lab it is surprising that so few studies relying on online samples appear in highly regarded journals.[3]

Yet, this does not necessarily mean that AMT responses are high quality. Experimental effects would potentially replicate if responses in both samples were of poor quality. AMT workers have a financial incentive to finish quickly, but respondents to a more traditional survey also would want to finish quickly. As part of an effort to improve the quality of responses, experimental economists ask respondents to participate in tasks with real stakes in order to measure underlying attitudes. For example, one could measure intergroup affect with survey questions, but measuring affect via a trust or dictator games avoids issues of social desirability (Fowler and Kam, 2007; Carlin and Love, 2013). By offering real stakes to answer a question, researchers can incentivize careful responding. An AMT worker, who answers the real stakes question too quickly, risks lower earnings which may defeat the purpose of quick responding. However, one could argue that the cost of this behavior is mitigated by the relatively low stakes. Hence, to demonstrate that AMT workers provide quality data, we need to demonstrate both that the responses are consistent across time, but also that the measures obtained via survey questions are consistent with those from a

[3]As of November 12, 2018, a simple google scholar search of the keyword "Amazon Mechanical Turk" reveals that since 2011, the year Horton et al. (2011) was published, *Experimental Economics*, *Games and Economic Behavior*, and *The American Economic Review* have published 10, 2 and 13 papers mentiong Amazon Mechanical Turk, respectably.

real (but low) stakes task.

While researchers are probably most often concerned with how workers respond within an experiment, there is also the worry that the workers are not who they say they are (e.g., a husband and wife share an AMT worker account) and/or have responses that vary substantially over time (possibly due to not paying attention to the questions or deliberately responding randomly); both implying responses would be more likely to be inconsistent. To address this concern, researchers have also explored the consistency (i.e., test-retest reliability) of online workers' responses. For example, Buhrmester et al. (2011) explores the test-retest reliability of AMT workers and finds that AMT workers consistently report their political beliefs and personality traits. Similar results are discussed in Rand (2012), which explores the consistency of reported age, gender, level of education, income, and belief in God; Holden et al. (2013), which explores the consistency of personality traits; and Shapiro et al. (2013), which explores the consistency of reported mental health. Such consistency may be considered surprising due to the lack of control but, in retrospect, less so because (Hauser and Schwarz, 2016) find many AMT workers perform better on attention checks than laboratory subjects.[4] If online workers are highly attentive, and consequently not responding randomly, it makes sense that their responses are also highly consistent.

## 3 Data

We ask two central questions: 1) is the data provided by AMT workers consistent? and 2) if so, are workers' responses economically meaningful – i.e., do their responses to survey questions correlate with their decisions in a real stakes task? The data we use is from eleven different experiments which were run from 2012 to 2018. To avoid selection bias, we use data from all of the experiments author one has run on AMT that asked workers to report

[4]Though in (Hauser and Schwarz, 2016), participation was restricted to, compared to our study, high quality workers (i.e., US workers with greater than 95% approval rating and more than 100 jobs approved).

at least two of the measures of interest (discussed below) and have access to their worker identification number (workerid) and the date the experiment was posted.[5] The dates of the experiments and the number of observations in each experiment are found in Table E.1 in the Appendix.[6] The timing of each of the variables of interest (e.g., age and impulsiveness test) varies widely across experiments. For example, in some experiments workers were asked about their willingness to take risk prior to the experiment (to conceal a treatment assignment variable) and, in others, the question occurred after the main experiment.

In general, the worker sample analyzed here is, in all likelihood, of lower quality than the samples used in Hauser and Schwarz (2016) and Arechar et al. (2018). This is because author one generally has little to no restrictions placed on who is eligible to participate in his experiments. What this means is that workers who have never completed an experiment, do not live in the US, and/or have a low approval rating are all usually eligible to participate.[7] This is important to highlight because Peer et al. (2014) demonstrate that high quality workers (i.e., those with approval ratings greater than 95%) are less likely to fail attention checks than low quality workers (i.e., those with approval ratings less than 95%). Thus, it is reasonable to suspect a higher percent of workers in the sample analyzed here, relative to the sample discussed in Hauser and Schwarz (2016) and Arechar et al. (2018) are randomly

[5]A description of each experiment is found in Appendix section 5. All experimental data is posted online. We will update the data set as we run more experiments.

[6]All of the experiments were written in html and javascript which was then copied and pasted into the area where HITs are edited. Each experiment consisted of multiple pages and workers could only continue to the next page after they had answered all of the questions on a given page. The incentives present in each of the experiments are low by laboratory standards but are probably on the high side for AMT - evidenced by author one's pay rating on turkopticon (a site where AMT workers grade requestors).

[7]Author one makes this design choice for two reasons. First, IP addresses (how worker location is identified) can be faked so there is no guarantee that the worker is even living in the United States. Second, for ethical reasons, author one believes that all workers should be given the opportunity to participate. Further, laboratory experiments do not limit participation to subjects who have completed a set number of experiments, so it seems contradictory to require this in the online setting.

entering text. In other words, the deck is stacked against us.

### 3.1 Questions

In assessing the consistency of workers' responses we evaluate four variables. The first two are basic demographic variables: age and gender. These measures provide only a low-bar test as answering those questions quickly and carelessly would take about as long as answering the questions accurately. It will serve to demonstrate that AMT workers are not "trolling" researchers (Lopez and Hillygus, 2018).

We perform more comprehensive analyses of two respondent personality traits: impulsiveness and willingness to take risk. Impulsiveness is a commonly studied trait in the psychology of criminals (Farrington, 1998) and has been used in political science to explain differences between liberals and conservatives (McAdams et al., 2013), partisan strength (Hatemi et al., 2009) and political violence (McDermott et al., 2013). Impulsiveness is measured using the Barratt Impulsivity Test (Stanford et al., 2009). Respondents indicate how often they engage in thirty different activities (e.g., "I plan tasks carefully") using a four point scale ranging from "Rarely/Never" (1) to "Almost Always/Always" (4). A worker's impulsiveness score is the sum of their responses.[8] To demonstrate that workers consistently report their level of impulsiveness, we use both workers' responses to each of the impulsiveness questions as well as their total score on the instrument.

Willingness to take risk is measured using the general risk attitudes question from the German Socio-Economic Panel Survey (SOEP) which was popularized in Dohmen et al. (2011). The English translation of the question is as follows:

> *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?*

[8] Some of these questions are reverse coded.

Workers answer this question on an 11 point scale ranging from 0 to 10 - with 0 corresponding to “I avoid risk" and 10 corresponding to “Fully prepared to take risks.”

Additionally, in two of the experiments respondents were asked to choose to play one lottery from a list of several lotteries; this is a common way to measure an individual’s risk profile in experimental economics (e.g., Holt and Laury, 2002; Dave et al., 2010). We will explain the details of the real stakes task later when we examine how it correlates with the answer to the SOEP risk attitudes question.

A convenient feature of the data sets generated using AMT is that they not only include workers’ decisions and responses but also a unique randomly generated worker identification number (workerid) and the date at which the HIT was posted online. We use the workerid to link the responses of workers who participated in at least two of the experiments presented in Table E.1 in the Appendix. The date is used to establish the order of the responses – allowing us to compare a respondent’s first response to their future responses.

### 3.2 AMT Workers

In the eleven experiments, there are 5,347 observations. 3,566 workers completed a single experiment. 730 workers completed at least two, and 223 workers completed three or more experiments. The average number of experiments completed by workers is 1.24 and the most experiments completed by a worker is eight. We did not contact workers who participated in one study to participate in future studies. Hence, the fact that respondents appear in this study in a panel nature is purely by chance. 4,127 workers reported their gender, 2,829 reported their age, 3,811 reported their willingness to take risk, and 2,730 completed the Barratt Impulsivity Test.[9]

Table 1 presents the summary statistics of the workers who participated in the exper-

[9]A breakdown of average responses by the day of the week the batch of the HIT is posted is found in section 1 of the Appendix. Detailed analysis of “day of the week” effects are not the primary purpose of this paper but is available upon request.

iments.[10] Each variable in Table 1 contains subscripts indicating a group and a response time. The first subscript corresponds to the order of the workers' response (i.e., 1 if first; 2 if last). Workers are classified into one of two groups. The first group (group 1) are workers who participated in one experiment. The second group (group 2) are workers who participated in two or more experiments. The group is indicated by the second subscript (i.e., 1 if group 1; 2 if group 2). For example, $\text{Risk}_{1,2}$ is the average earliest reported willingness to take risk for workers who completed more than one experiment.

The distributions of Age, Risk, and IMP by group, and order of response are found in Figure 4. Figure 4 and Table 1 suggest that there are differences in the workers who completed more than one experiment compared to those who only completed a single experiment. Workers who completed more than one experiment are significantly more likely to report being male (test of proportions: 0.561 vs 0.628, $p = 0.001$), are less impulsive (t-test: 60.623 vs 58.594, $p < 0.001$; u-test: $p < 0.001$), are less willing to take risk (t-test: 5.364 vs 5.037, $p = 0.005$; u-test: $p = 0.005$), and older (t-test: 32.149 vs 33.856, $p < 0.001$; u-test: $p < 0.001$) than workers who completed a single experiment.[11] This means the experienced AMT participants have somewhat different characteristics than those who are less experienced, which may lead to questions regarding external validity relating to the consistency of responses. Yet, it does not affect our ability to test whether experienced and inexperienced workers provide economically meaningful responses.[12]

[10] Careful readers will note that the sum of the variables in Table 1 do not add up to the "correct" total. This is due to the fact that workers could have completed experiments that did not ask them to report one or two of the measures. So, for example, while a worker could have completed more than one experiment, she could have actually only reported her age once.

[11] All t-tests assume unequal variance.

[12] Note that the averages presented in Table 1 do not match the averages and proportions reported when we are comparing responses across groups. This is not a mistake but reflects the fact not all HITs asked the same questions. See Footnote 10 for a more concrete example.

Table 1: Summary Statistics

| Responded Once | | | | | |
|---|---|---|---|---|---|
| Variable | Obs | Mean | Std. | Min | Max |
| $Male_{1,1}$ | 3398 | 0.561 | 0.496 | 0 | 1 |
| $Age_{1,1}$ | 2136 | 32.149 | 10.028 | 6 | 69 |
| $Risk_{1,1}$ | 3090 | 5.364 | 2.691 | 0 | 10 |
| $IMP_{1,1}$ | 2055 | 60.623 | 11.628 | 30 | 110 |
| Responded More than Once (First Response) | | | | | |
| Variable | Obs | Mean | Std. | Min | Max |
| $Male_{1,2}$ | 678 | 0.645 | 0.479 | 0 | 1 |
| $Age_{1,2}$ | 418 | 33.117 | 9.536 | 18 | 67 |
| $Risk_{1,2}$ | 553 | 4.911 | 2.793 | 0 | 10 |
| $IMP_{1,2}$ | 406 | 58.628 | 11.106 | 32 | 92 |
| Responded More than Once (Last Response) | | | | | |
| Variable | Obs | Mean | Std. | Min | Max |
| $Male_{2,2}$ | 678 | 0.639 | 0.481 | 0 | 1 |
| $Age_{2,2}$ | 418 | 33.969 | 9.663 | 19 | 69 |
| $Risk_{2,2}$ | 553 | 4.929 | 2.824 | 0 | 10 |
| $IMP_{2,2}$ | 406 | 58.369 | 12.441 | 31 | 120 |

*Summary statistics of variables of interest by group and order of response. Male is the reported gender (1 if male; 0 if female). IMP is Barratt Impulsiveness Test score. Risk is general willingness to take risk. Age is reported age in years. First subscript number indicates whether this is for a respondent's first (1) or most recent response (2). Second subscript number indicates whether the respondent participated in only one survey (1) or at least two experiments (2).*

In the coming analysis, we will primarily consider the respondents who participated in more than one study as those are the only respondents who we can show are consistent in their responses. However, we will later use some participants who participated only once to compare their responses to the risk question to their decisions made in the real stakes lottery task. As we will demonstrate, the relationship between subjective risk preferences and choices made in the incentivized lottery experiment, for workers who participated in only one experiment, is not qualitatively different (e.g., similar size and significance) than the relationship observed in workers who participated in more than one experiment.

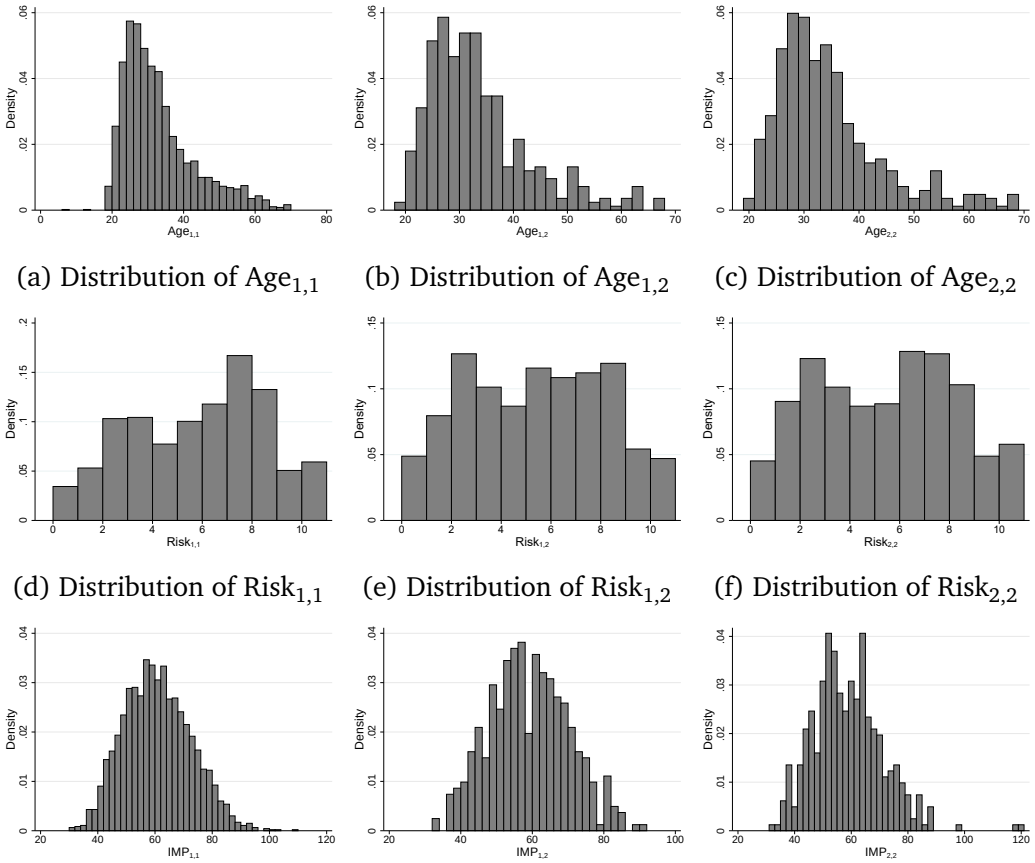
## 4 Results

### 4.1 Age and Gender Consistency

As a "low-bar" test of the consistency of responses, we start by exploring workers' first reported gender and age and compare these responses to their final responses. The average number of days between first reported gender(age) and last reported gender(age) is 401(289). Overall, 666 workers consistently reported their gender. 237 of these workers reported being female and 429 of these workers reported being male. Eight(four) workers first reported being male(female) but reported being female(male) in the last experiment they completed. Thus, workers consistently report their gender 98.23 % of the time – which is a higher gender consistency (96%) reported in Rand (2012).[13]

[13]Given that recent studies estimate that between 0.4% and 0.6% of the US population identifies as transgender (Flores et al., 2016), some of the observed inconsistency possibly represents real changes in gender identities rather than a mistake or careless response. Evidence in support of this can be seen when looking at the workers who participated in more than one experiment and changed their reported gender. For example, one worker, first reported being male and then subsequently reported being female in two later experiments. Obviously, it is impossible to say whether or not the first response given by this worker was a mistake or not but it is noteworthy to mention that this worker was very consistent in their reported willingness to take risk ($Risk_{1,2}$=4 and $Risk_{2,2}$=5) and impulsiveness ($IMP_{1,2}$=52 and $IMP_{2,2}$=49).

Figure 1: Distribution of workers' characteristics by group and completion order.




(a) Distribution of $\text{Age}_{1,1}$ (b) Distribution of $\text{Age}_{1,2}$ (c) Distribution of $\text{Age}_{2,2}$

(d) Distribution of $\text{Risk}_{1,1}$ (e) Distribution of $\text{Risk}_{1,2}$ (f) Distribution of $\text{Risk}_{2,2}$

(g) Distribution of $\text{IMP}_{1,1}$ (h) Distribution of $\text{IMP}_{1,2}$ (i) Distribution of $\text{IMP}_{2,2}$

*First row of sub-figures in Figure 4 corresponds to age, second row corresponds to reported willingness to take risk, and the third row is Impulsivity Test scores. First column are the distributions of characteristics of the workers who only participated once (group 1). Middle column are the distributions of characteristics reported in the first experiment among workers who participated in two or more of the experiments (group 2). Third column are the distributions of characteristics reported in the last experiment completed among those who completed more than one experiment (group 2).*

418 workers reported their age more than once. The correlation between first and last reported age is positive and significant ($r = 0.977$, $p < 0.0001$). The average reported age when workers completed their first(last) experiment 33.12(33.97) which translates into a difference of .85 years. While this difference is statistically significantly different from zero (t-test: $p < 0.001$) it is expected considering workers will age over the time between experiments. Yet, this difference is not statistically different from the difference in the time they last reported their age and first reported their age divided by 365 (t-test: 0.85 vs .793, $p = 0.174$). Just as with gender there is some inconsistency. There are eight "Benjamin Buttons" who got younger (i.e., their last reported age was less than first reported age) and five subjects who aged more than was reasonably possible over the six year period.[14] Nonetheless, of the workers who reported their age more than once, roughly 91% of workers who reported a final age that was within 1 year of their first reported age plus the number of days between the two reported ages divided by 365. While this consistency is lower than the age consistency (93%) reported in Rand (2012), it is not, in our opinion, meaningfully different. Overall, these results suggest, for simple demographic questions at least, AMT respondents are fairly consistent.

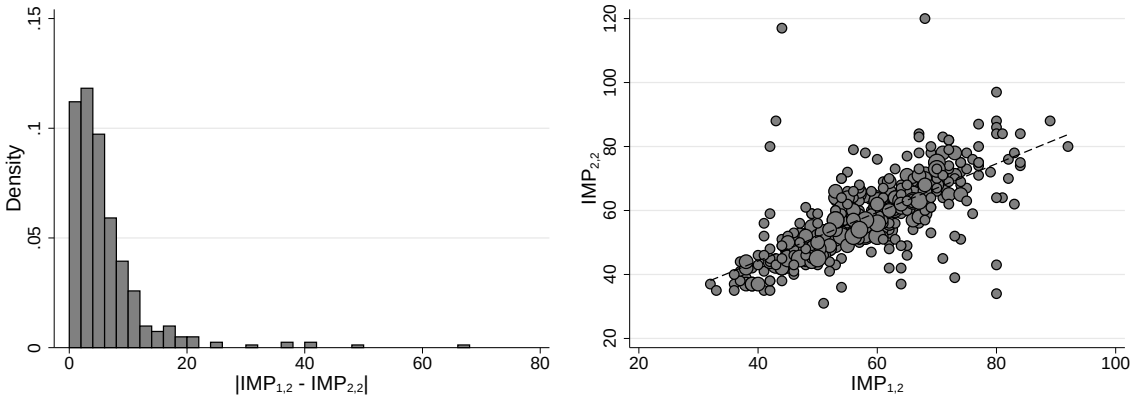
### 4.2 Consistency in Impulsivity

We now present results suggesting that workers are consistently reporting their impulsivity. The distribution of the absolute difference between workers' first and last Impulsiveness Test score is found in Figure 2a. In Figure 2b, we present a scatter plot of $\text{IMP}_{1,2}$ and $\text{IMP}_{2,2}$. The average time difference between workers' first and last Impulsivity Test is 339 days. As with the previous demographics and characteristics, workers consistently report their impulsivity.

The difference in workers' first and last Impulsivity Test score is not statistically dif-

[14] For example, one worker aged 17 years over 163 days.

Figure 2: Distribution of the Absolute Differences and Scatter Plots of First and Last Impulsivity Test Responses.




(a) Distribution of $|IMP_{1,2}\text{-}IMP_{2,2}|$ (b) Scatter Plot of $IMP_{1,2}$ and $IMP_{2,2}$

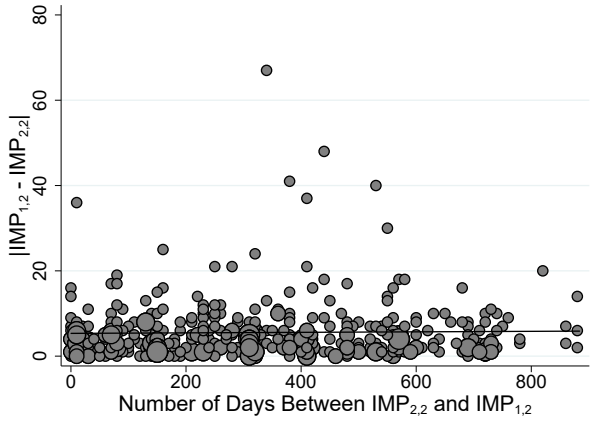
*Figure 2a presents the distribution of the absolute difference of workers' first Impulsiveness Test score and final Impulsiveness Test score. Figure 2b presents a scatter plot of workers' first Impulsiveness Test score ($IMP_{1,2}$) and final reported willingness to take risk ($IMP_{2,2}$). X-axis corresponds to first reported willingness to take risk/test score while y-axis is final reported willingness to take risk/test score. Dashed lines are fitted regression lines. Dot sizes, in the scatter plots, indicate the proportion of responses.*

ferent (paired t-test: 58.628 vs 58.369, $p = 0.5843$). While the two scores are highly correlated ($r = 0.679$, $p < 0.001$) the observed correlation is not as high as what is reported in Stanford et al. (2009) who reports a one month correlation across test scores of 0.83 using a sample made up of college students and subjects recruited from the cities of Winston-Salem, NC and Houston, TX. As can be seen in Figure 2a, this difference is being driven by a few very inconsistent workers. For example, if we remove the eight most inconsistent workers (or less than 2 % of the sample of workers who completed the Impulsivity Test more than once), $r$ increases to .804.[15]

The number of days between measures does not affect the consistency of the responses. This can be seen in Figure 3 which plots the absolute difference between first

[15]Interestingly, one of these workers inconsistently reported their gender which might suggest this an extreme troll.

Figure 3: Difference in Impulsivity Test Responses Over Time.

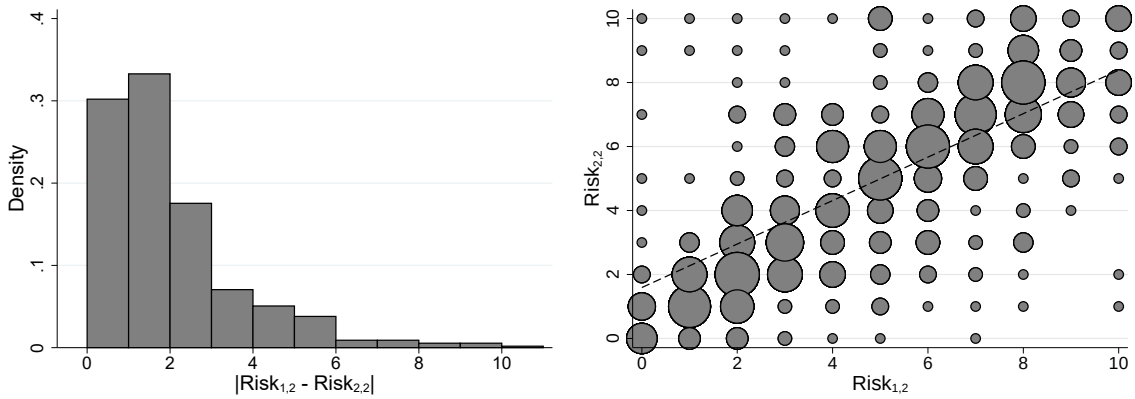



*The solid lines is the fitted regression line. Dot sizes indicate the proportion of responses.*

and last scores on the impulsivity measure on the y-axis and the number of days between responses on the x-axis. When we estimate an OLS model using the absolute difference as a dependent variable and the number of days between responses as the independent variable, we find the coefficient on the number of days between measures is small and not statistically significant ($coef = 0.001$, $p = 0.594$, $n = 406$).

In Appendix section 4, we present the result for each of thirty items in test. The difference between initial and final responses to one item has a p-value less than 0.05 with another two items having p-values less than 0.1. Given the multiple tests, we should adjust our p-values for multiple comparisons. Whether we account for "false discovery rate" – the expected proportion of false positives (Type I errors) (Benjamini and Hochberg, 1995) – or "Family-Wise Error Rate" – the probability of incorrectly rejecting even one null hypothesis – using the common Bonferroni correction, there are no statistically significant differences between the first and final responses to any of the items.

Figure 4: Distribution of the Absolute Differences and Scatter Plots of First and Last Impulsivity Test Responses.




(a) Distribution of $|\text{Risk}_{1,2}\text{-Risk}_{2,2}|$

(b) Scatter Plot of $\text{Risk}_{1,2}$ and $\text{Risk}_{2,2}$

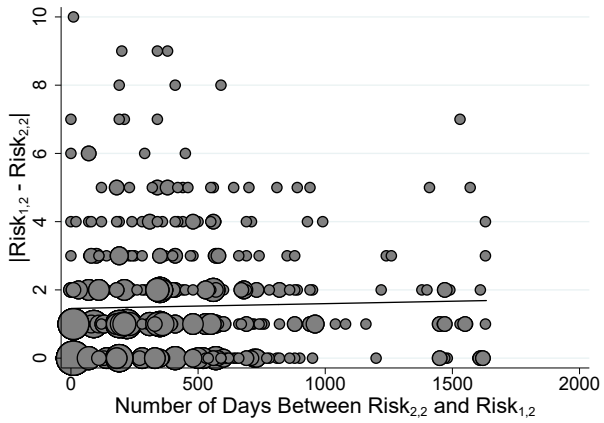
*Figure 4a presents the distribution of the absolute difference of workers' first reported general willingness to take risk and final reported general willingness to take risk. Figure 4b presents a scatter plot of workers' first reported willingness to take risk ($Risk_{1,2}$) and final reported willingness to take risk ($Risk_{2,2}$). Dashed lines are fitted regression lines. Dot sizes, in the scatter plots, indicate the proportion of responses.*

### 4.3 Consistency in Willingness to Take Risk

The distribution of the absolute difference between workers' first and last reported willingness to take risk is found in Figure 4a. In Figure 4b, we present a scatter plot of $\text{Risk}_{1,2}$ and $\text{Risk}_{2,2}$. The average time difference between workers' first and last response to the risk question is 418 days. The difference between reported willingness to take risk the first and last time they completed an experiment is not statistically significant (paired t-test: 4.911 vs 4.929, $p = 0.852$) and, as expected, the responses themselves are highly correlated ($r = 0.672$, $p < 0.001$). Further, the reported reliability (i.e., test and re-test correlation) is higher than the 30-49 day reliability ($r = 0.60$) reported in the SOEP manual (Richter et al., 2017).

Additionally, we once again see no evidence that the difference in reported willingness to take risk increases in the number of days between the first and last time it is reported. We

Figure 5: Difference in Willing to Take Risk Over Time.




*The solid lines is the fitted regression line. Dot sizes indicate the proportion of responses.*

show this in Figure 5 which plots the absolute difference between first and last responses to risk question on the y-axis and the number of days between responses on the x-axis. When estimating the absolute difference between first and last response, with OLS, using the number of days between responses as the independent variable, we find the coefficient on the number of days between measures to be small and not statistically significant ($coef =$ 0.0001, $p = 0.472$, $n = 553$).

### 4.3.1 Risk Measured with Real Stakes

We have shown that AMT respondents provide consistent responses, but this does not necessarily mean the data is of high quality or economically meaningful. Consistent responses could be the result of careful consideration or a basic decision rule – for example, always answer the middle option. We now show that the data AMT workers provide is economically meaningful. We do so by showing that subjective risk preferences correlate in the expected direction with workers' decisions in a simple real stakes task: an incentivized

lottery experiment (Johnson and Webb, 2016; Gibson and Johnson, 2018).[16]

In the experiment, workers are given a set of 20 lotteries (shown in Table 2) that vary in the probability they will be successful and the payoff that will be paid if the lottery turns out to be successful. Lotteries that are riskier (i.e., lower in index number) have a higher payoff (if successful) while lotteries that are safer (i.e., higher in index number) have a lower payoff (if successful).[17] After being shown the possible lotteries that they can select, and each lottery's expected value, workers are asked to select the lottery that they wish to play for real stakes. The advantage of this simple risk task compared to some others available is that the measure does not conflate risk preference and math ability as is the case with some more complex measures, which makes it particularly suitable for the environment and population.

The stakes of the experiment are quite low relative to laboratory experiments. Workers, on average, earned about $1.15 for their decisions. 122 workers completed the lottery experiment. The average subjective willingness to take risk of workers, taken at the time of the incentivized experiment is 4.76. The correlation between these two variables is -0.332 and is statistically significant ($p = 0.0002$). The average index number of the lottery selected by these workers is 11.75 which is statistically significantly greater than the index number of the safer of two lotteries that are expected utility maximizing (lottery 11 in Table 2) assuming risk neutral preferences (t-test: $p = 0.039$). Overall, roughly 67% of subjects selected a lottery that is consistent with some degree of risk averse preferences. Of subjects who completed the lottery experiment and a prior experiment in which they were asked to report their general willingness to take risk, this figure rises to approximately 76%.

[16]Note here we are focusing on the control treatment of Johnson and Webb (2016) that had workers select a single lottery rather than a bundle of lotteries. Gibson and Johnson (2018) had workers only select a single lottery.

[17]In the actual experiment, lotteries are indexed by a letter of the alphabet rather than number.

Table 2: Lotteries Used in Real Stakes Task

| Lottery | Prob. | Prize | E.V. | $Obs_1$ | $Obs_2$ |
|---|---|---|---|---|---|
| 1 | 0.05 | $5.00 | $0.25 | 9 | 2 |
| 2 | 0.1 | $4.75 | $0.48 | 0 | 0 |
| 3 | 0.15 | $4.50 | $0.68 | 3 | 1 |
| 4 | 0.2 | $4.25 | $0.85 | 0 | 0 |
| 5 | 0.25 | $4.00 | $1.00 | 3 | 0 |
| 6 | 0.3 | $3.75 | $1.13 | 0 | 0 |
| 7 | 0.35 | $3.50 | $1.23 | 4 | 1 |
| 8 | 0.4 | $3.25 | $1.30 | 1 | 0 |
| 9 | 0.45 | $3.00 | $1.35 | 5 | 2 |
| 10 | 0.5 | $2.75 | $1.38 | 15 | 2 |
| 11 | 0.55 | $2.50 | $1.38 | 14 | 1 |
| 12 | 0.6 | $2.25 | $1.35 | 8 | 7 |
| 13 | 0.65 | $2.00 | $1.30 | 14 | 5 |
| 14 | 0.7 | $1.75 | $1.23 | 9 | 2 |
| 15 | 0.75 | $1.50 | $1.13 | 12 | 5 |
| 16 | 0.8 | $1.25 | $1.00 | 8 | 1 |
| 17 | 0.85 | $1.00 | $0.85 | 9 | 4 |
| 18 | 0.9 | $0.75 | $0.68 | 4 | 1 |
| 19 | 0.95 | $0.50 | $0.48 | 2 | 0 |
| 20 | 1 | $0.25 | $0.25 | 2 | 0 |

*Prob. is the probability the lottery will favorable to the subject, Prize is the amount won if the lottery is favorable, and E.V. is the expected value of the lottery.* $Obs_1$ *is the number of subjects who selected a given lottery.* $Obs_2$ *is the number of subjects who selected a given lottery and completed a prior experiment in which they were asked to report their general willingness to take risk.*

In both Johnson and Webb (2016) and Gibson and Johnson (2018) workers also reported their general willingness to take risk. 122 workers participated in either the control treatment of Johnson and Webb (2016) ($n = 47$) or Gibson and Johnson (2018) ($n = 75$). We now demonstrate that workers' subjective willingness to take risk correlates with their lottery selections. We do so in two ways in Table 3's tobit models.[18] Though we are aware of its shortcomings, the tobit is our preferred estimator because of censoring in the dependent variable and the ease of interpretation. In Models 1 and 2, both measures are taken at the same point in time. In Models 3 and 4, we use the 34 respondents who participated in more than one study. Moreover, instead of using their response to the subjective risk preference question from the same study as the lottery experiment, we use their subjective risk preference response from the first study in which they participated.[19] This allows us to determine if survey responses from months (even years!) earlier are related to their real stakes behavior. The average gap, in days, between workers' first reported willingness to take risk and the lottery experiment is 521 days – more than a year.[20] So these older measures are quite old relative to other studies that tend to use a 1-month reliability. Consequently, one could think of these measures as a lower bound in terms of relevancy as they correspond to an individual's willingness to take risk in the past.[21]

The subjective risk variable is measured with larger values indicating a greater willingness to take risk. At the same time, the riskier lottery choices are lower in index num-

[18] Similar results (in terms of significance and direction) are observed in all alternative models considered, i.e., ordered probit, interval regression (dependent variable is the risk parameter inferred from lottery choices assuming CRRA utility), and OLS. Results using alternative estimators are found in the Appendix.

[19] We do not use control variables in these models because of the smaller sample size.

[20] Roughly, 35% of subjects who completed the lottery experiment and a prior experiment in which they were asked to report their willingness to take risk completed both experiments within a year. 20% of subjects completed the lottery experiment more than 2 years after the first experiment in which they were asked to report their willingness to take risk.

[21] So not only are we using data of dubious quality but *old* data of dubious quality.

Table 3: Subjective 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 |
| Male | | | 1.452 | 0.881 | | | | |
| Age | | | 0.041 | 0.043 | | | | |
| Impulsivity | | | -0.003 | 0.036 | | | | |
| Constant | 14.404 | 0.835 | 12.319 | 2.817 | 14.333 | 2.294 | 14.678 | 3.751 |
| Obs | 122 | | 119 | | 34 | | 34 | |
| $R^2$ | 0.020 | | 0.024 | | 0.059 | | 0.059 | |
| LL | 9 | | 8 | | 2 | | 2 | |
| UL | 2 | | 2 | | 0 | | 0 | |

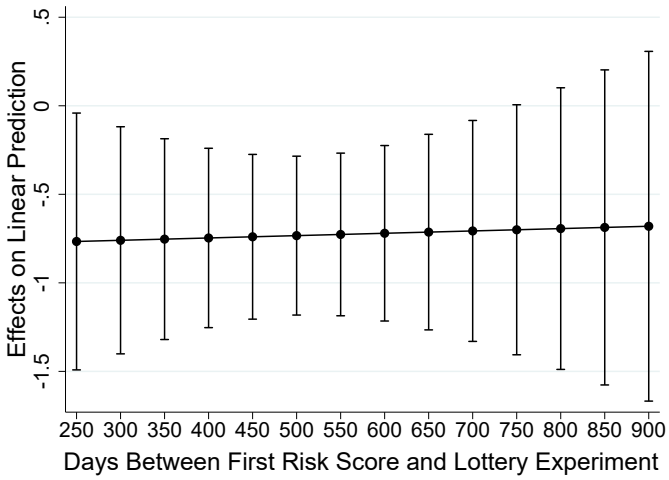
*Tobit models. The dependent variable is the index number of lottery selected: larger values indicate less risky choices. LL(UL) indicates the number of workers who selected lottery 1(20).*

ber. Therefore, since we expect subjective risk preferences to correlate with decisions made when there are real stakes, we expect to observe an inverse correlation between workers' response to the risk question and the lottery selected. This is exactly what is observed in Model 1 and Model 2 of Table 3. This suggests that the responses to the risk measure are not only consistent, but are economically meaningful. Hence, the data acquired via AMT worker responses are, on average, high quality measures.

In Models 3 and 4, we go a step further. First, as Model 3 shows, the same expected negative correlation holds – even if the subjective response was from an earlier time point. Second, as Model 4 shows, the size of correlation does not differ as the number of days since the first subjective response increases. We demonstrate this by including an interaction effect between the subjective response and the days since the response was given (Figure 6).

As Figure 6 shows, the marginal effect of subjective risk preferences on an individual's

Figure 6: Marginal Effect of Willingness to Take Risk as Number of Days Since 1st Response Increases.




*Marginal effect estimated from Model 4 in Table 3. Bars represent 95% confidence intervals.*

lottery choice is always around -0.7. The relationship is statistically significant at the .05 level for responses given up to 2 years prior. After that point, we simply do not have enough observations to show the correlation with any confidence.

Naturally, it may be that the statistically significant relationship between subjective risk preferences and decisions made in the incentivized lottery experiment is being driven only by workers who participated in more than one experiment. This would be troublesome because it could suggest that only the decisions made by workers who participate in more than one experiment are economically relevant. Given that workers who participate in more than one experiment tend to be older, less willing take risk, and less impulsive - this seems like a reasonable possibility. To explore whether or not this is the case, we re-estimate Models 1 and 2 from Table 3 using data from workers who participated in one experiment (EXP=1) and workers who participate in more than one experiment (EXP>1). These results are presented in Table 4. As can be seen in Table 4, the direction and magnitude of the coefficient estimates are similar to what is found in Table 3. The results presented in Table 4

therefore suggest that the subjective risk preferences of workers who participate in only one experiment predict risky decision making in the real stakes experiment. Further, coefficients in the models that use data from workers who completed only a single experiment are in the same ballpark as counterparts presented in Table 3 as well as models that use data from workers who completed more than one experiment.[22]

Table 4: Subjective Risk Preferences and Risky Decision Making in a Real Stakes Task (Group 1 vs Group 2)

| | Model 1 EXP=1 | | Model 2 EXP>1 | | Model 3 EXP=1 | | Model 4 EXP>1 | |
|---|---|---|---|---|---|---|---|---|
| | Coef. | S.E | Coef. | S.E | Coef. | S.E | Coef. | S.E |
| Contemporaneous Risk | -0.573 | 0.192 | -0.605 | 0.242 | -0.493 | 0.207 | -0.588 | 0.239 |
| Male | | | | | 1.077 | 1.184 | 2.175 | 1.240 |
| Age | | | | | 0.006 | 0.057 | 0.096 | 0.061 |
| Impulsivity | | | | | -0.059 | 0.050 | 0.076 | 0.053 |
| Constant | 14.756 | 1.093 | 14.025 | 1.291 | 17.426 | 3.565 | 4.765 | 4.395 |
| Obs | 69 | | 53 | | 66 | | 53 | |
| $R^2$ | 0.0211 | | 0.0194 | | 0.0257 | | 0.0393 | |
| LL | 5 | | 4 | | 4 | | 4 | |
| UL | 1 | | 1 | | 1 | | 1 | |

*Tobit models. The dependent variable is the index number of lottery selected: larger values indicate less risky choices. LL(UL) indicates the number of workers who selected lottery 1(20).*

In all, the results in this section are an important indication that AMT workers give not only consistent responses, but responses that map onto actual behavior. Further, because the responses are consistent, it does not matter when the response was recorded. Responses taken several months prior are just as good at predicting risky behavior in a real stakes task as responses taken at the same time. This suggests that, on average, AMT workers provide quality data – at least for the questions we asked respondents.

[22]As before, similar results (in terms of significance and direction) are observed in all alternative models considered, i.e., ordered probit, interval regression (dependent variable is the risk parameter inferred from lottery choices assuming CRRA utility), and OLS. Results using alternative estimators are found in the Appendix.

## 5 Discussion and Conclusion

Critics of studies using Amazon Mechanical Turk question the validity of studies using the platform for a wide variety of reasons. Many of these criticisms are valid. Namely, experiments involving an element of social interaction or those that require specific subject pools/samples are probably not best suited for the platform. However, in many experiments, these types of criticism are not applicable. Therefore, the primary concerns become those related to consistency and economic relevancy of workers' responses (i.e., is the data good?). We show that workers on Amazon Mechanical Turk consistently report their basic demographics, subjective risk preferences, and impulsivity. Further, their responses are economically relevant in, at the very least, a basic lottery experiment - though the preponderance of evidence suggests relevancy far beyond a basic lottery experiment. Subjective willingness to take risk is significantly correlated with behavior in an incentivized lottery experiment and in the direction one would expect. In sum, the results suggest the quality of data obtained via AMT is not significantly harmed by the lack of control over the conditions under which the responses are recorded or the stakes.

Given that the stakes on AMT are very low relative to the lab (c.f., Horton et al., 2011; Amir et al., 2012) and the fact that, unlike the lab, workers are not monitored, it is natural to wonder why the decisions made by AMT workers are not all that different from the decisions made by laboratory subjects and why their responses have high test-retest reliability. As a possible explanation, we would like to remind readers that poor quality responses by workers can lead to rather severe penalties. For example, if a worker submits a HIT and does not answer any of the questions, the requestor can reject the submission and the worker will be paid nothing. Additionally, requestors have the option of blocking workers which will prevent the worker from completing any of the requestor's HITs in the future. Rejections and blocks themselves are also quite costly to workers. If a worker receives too many blocks, they risk having their account deleted. Furthermore, each worker

has an approval rate, which is determined by the percent of HITs the worker has completed that have not been rejected. If a worker's approval rate gets too low, it reduces the number of HITs they are eligible to complete. This is because the worker's approval rate is one of the "canned" qualifications that is available to all requestors.[23] These punishments can be very costly to worker – especially so when one considers the fact that 25% percent of workers on AMT report that "all" or "most" of their income comes from AMT.[24]

Researchers need to, however, use Amazon Mechanical Turk with care. As Arechar et al. (2017) shows, the population completing tasks on Amazon Mechanical Turk tends to be more male during the weekdays and the studies used to generate the data set that is analyzed here support this finding. Researchers using AMT should provide the time the HIT was posted along with any observed differences in their samples from ongoing large scale projects. Further, while the population of workers is generally more representative of the US population than typical university subject pools, there are still significant differences between the worker population and US population. Researchers need to take these differences seriously before making external validity claims. Additionally, recently there has been concern raised in informal academic circles (e.g., Twitter) that computer programs (called "bots") are being used to complete HITs rather than humans. Evidence in support mechanized survey responses is based off of suspicious geocodes (inferred from IP addresses). While we do not dispute that bots may pose a problem, our work here suggests that at the current moment, these concerns are more molehills than mountains. This is because if bots are a problem we would not expect subjective willingness to take risk to be correlated with decisions made in the incentivized lottery experiments. However, this

[23]For example, if the requestor does not want workers, with approval rates less than 90 percent, to participate, they can tell AMT that only workers with an approval rate greater than 90 percent can complete the HIT.

[24]Paul Hitlin "Research in the Crowdsourcing Age, a Case Study" *http://www.pewinternet.org* July 7, 2016. July 24, 2017.

conjecture needs to be taken with a grain of salt and may only be relevant to our study.[25] In terms of future steps, normatively, research using AMT is lacking in standardization. Unlike the lab, there are few established norms for conducting research on AMT or online, generally. A consensus of best practices (e.g., pay, reasons for blocking workers, and bot prevention tactics) needs to be reached for comparisons across studies to be meaningful. Such standardization could also foster increased experimental control.

To close, there are three caveats regarding the results presented in this manuscript. First, this paper is not about the overall demographics of the AMT worker population. Worker demographics vary across the day and week and there is evidence of long term trends in changing demographics. Given that the experiments were posted during relatively consistent times of day and over a long period of time, there is no reason to believe that the demographics presented will match the demographics of today's AMT worker population. Indeed, they do not. Second, this paper is not about identifying the inter-temporal correlation of various worker characteristics and demographics under the most controlled conditions allowed in the environment. The experiments from which we include data vary widely and the order of the measures taken in the experiments do as well. Further, we try to keep as many observations as possible. This means that we include data from workers who provided inconsistent answers, did not complete the experiment, or failed an English comprehension/attention check question.[26] Third, it is well-known that the population of AMT workers may not be suitable for all research questions. While it is true that, demographically, AMT workers are significantly more varied than university subject pools (Ipeirotis, 2010; Buhrmester et al., 2011; Behrend et al., 2011), the AMT population is not represen-

[25]The experiments analyzed in this study all had, multiple pages, and are somewhat old. Further, the HITs themselves were programmed in notepad using a mixture of javascript and HTML and did not link to an external site (e.g., surveymonkey) where subjects completed the survey.

[26]So in some sense, the data analyzed in this study is the experimentalist's worst nightmare.

tative of the US population as a whole. American workers on AMT generally have a lower reported income than the overall population of the US (c.f., Paolacci et al., 2010; Berinsky et al., 2012; Levay et al., 2016), are more likely to report to be unemployed (e.g., Ipeirotis, 2010; Goodman et al., 2016), younger, and are more ideologically liberal (Berinsky et al., 2012).[27] Thus, this paper is not arguing that AMT is a suitable tool for all research questions. Instead, as we have shown, the purpose of this paper is to demonstrate that even in inconsistent settings, with low stakes, in an uncontrolled environment, with some of the worst workers possible, workers on AMT can provide consistent and economically meaningful data. In sum, given this research and the abundance of prior research discussed in the introduction, it is difficult to argue the merit of criticizing online studies on the basis of low stakes and lack of control.

[27]However, AMT workers are fairly similar to samples created using other online survey platforms (Huff and Tingley, 2015).

**References**


Amir, Ofra, David G Rand, and Kobi Gal (2012) 'Economic Games on the Internet: The Effect of $1 Stakes.' *PloS one* 7(2), e31461

Arechar, Antonio A, Gordon T Kraft-Todd, and David G Rand (2017) 'Turking Overtime: How Participant Characteristics and Behavior Vary Over Time and Day on Amazon Mechanical Turk.' *Journal of the Economic Science Association* 3(1), 1–11

Arechar, Antonio A, Simon Gächter, and Lucas Molleman (2018) 'Conducting Interactive Experiments Online.' *Experimental Economics* 21(1), 99–131

Behrend, Tara S, David J Sharek, Adam W Meade, and Eric N Wiebe (2011) 'The Viability of Crowdsourcing for Survey Research.' *Behavior Research Methods* 43(3), 800

Benjamini, Yoav, and Yosef Hochberg (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.' *Journal of the Royal Statistical Society, Series B (Methodological)* 57(1), 289–300

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz (2012) 'Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk.' *Political Analysis* 20(3), 351–368.

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling (2011) 'Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?' *Perspectives on Psychological Science* 6(1), 3–5

Carlin, Ryan E., and Gregory J. Love (2013) 'The Politics of Interpersonal Trust and Reciprocity: An Experimental Approach.' *Political Behavior* 35(1), 43–63

Clifford, Scott (2014) 'Linking Issue Stances and Trait Inferences: a Theory of Moral Exemplification.' *The Journal of Politics* 76(3), 698–710

Clifford, Scott, and Ben Gaskins (2016) 'Trust Me, I Believe in God: Candidate Religiousness as a Signal of Trustworthiness.' *American Politics Research* 44(6), 1066–1097

Converse, Philip E. 'Information Flow and the Stability of Partisan Attitudes.' *Public Opinion Quarterly* 26(4), 578–599

Dave, Chetan, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas (2010) 'Eliciting Risk Preferences: When is Simple Better?' *Journal of Risk and Uncertainty* 41(3), 219–243

Del Ponte, Alessandro, Andrew W. Delton, Reuben Kline, and Nicholas A. Seltzer (2017) 'Passing It Along: Experiments on Creating the Negative Externalities of Climate Change.' *Journal of Politics* 79(4), 1444–1448

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner (2011) 'Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences.' *Journal of the European Economic Association* 9(3), 522–550.

Farrington, David P. (1998) 'Predictors, Causes, and Correlates of Male Youth Violence.' *Crime and Justice* 24, 421–475

Flores, Andrew R, Taylor NT Brown, and Jody Herman (2016) *Race and Ethnicity of Adults Who Identify as Transgender in the United States* (Williams Institute, UCLA School of Law)

Fowler, James H, and Cindy D Kam (2007) 'Beyond the self: Social identity, altruism, and political participation.' *The Journal of politics* 69(3), 813–827

Gibson, John, and David Johnson (2018) 'Assessing the Stability of Risk Preferences Online.' Technical Report, University of Central Missouri

Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema (2013) 'Data Collection in

a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples.' *Journal of Behavioral Decision Making* 26(3), 213–224

Goodman, William K, Ashley M Geiger, and Jutta M Wolf (2016) 'Differential Links between Leisure Activities and Depressive Symptoms in Unemployed Individuals.' *Journal of Clinical Psychology* 72(1), 70–78.

Hatemi, Peter K., John R. Alford, John R. Hibbing, Nicholas G. Martin, and Lindon J. Eaves (2009) 'Is There a 'Party' in Your Genes?' *Political Psychology* (3), 584–600

Hauser, David J, and Norbert Schwarz (2016) 'Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than do Subject Pool Participants.' *Behavior Research Methods* 48(1), 400–407.

Holden, Christopher J, Trevor Dennie, and Adam D Hicks (2013) 'Assessing the Reliability of the M5-120 on Amazon's Mechanical Turk.' *Computers in Human Behavior* 29(4), 1749–1754

Holt, Charles A, and Susan K Laury (2002) 'Risk Aversion and Incentive Effects.' *American Economic Review* 92(5), 1644–1655.

Horton, John J, David G Rand, and Richard J Zeckhauser (2011) 'The Online Laboratory: Conducting Experiments in a Real Labor Market.' *Experimental Economics* 14(3), 399–425

Huff, Connor, and Dustin Tingley (2015) '"Who are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents.' *Research & Politics* 2(3), 2053168015604648

Ipeirotis, Panagiotis G (2010) 'Demographics of Mechanical turk.'

Johnson, David B, and Matthew D Webb (2016) 'Decision Making with Risky, Rival Outcomes: Theory and Evidence.' Technical Report, Carleton University, Department of Economics

Krupnikov, Yanna, and Adam Seth Levine (2014) 'Cross-Sample Comparisons and External Validity.' *Journal of Experimental Political Science* 1(1), 59–80

Levay, Kevin E, Jeremy Freese, and James N Druckman (2016) 'The Demographic and Political Composition of Mechanical Turk Samples.' *SAGE Open* 6(1), 2158244016636433.

Lopez, Jesse, and D. Sunshine Hillygus (2018) 'Why So Serious?: Survey Trolls and Misinformation.' Available at SSRN: https://ssrn.com/abstract=3131087 or http://dx.doi.org/10.2139/ssrn.3131087

McAdams, Dan P., Kathrin J. Hanek, and Joseph G. Dadabo (2013) 'Themes of Self-Regulation and Self-Exploration in the Life Stories of Religious American Conservatives and Liberals.' *Political Pyschology* 34(2), 201–219

McDermott, Rose, Chris Dawes, Elizabeth Prom-Wormley, Lindon Eaves, and Peter K. Hatemi (2013) 'MAOA and Aggression: A Gene–Environment Interaction in Two Populations.' *The Journal of Conflict Resolution* 57(6), 1043–1064

Milita, Kerri, Elizabeth N. Simas, John Barry Ryan, and Yanna Krupnikov (2017) 'The Effects of Ambiguous Rhetoric in Congressional Elections.' *Electoral Studies* 46(April), 48–63

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese (2015) 'The Generalizability of Survey Experiments.' *Journal of Experimental Political Science* 2(2), 109–138

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis (2010) 'Running Experiments on Amazon Mechanical Turk.'

Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti (2014) 'Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk.' *Behavior Research Methods* 46(4), 1023–1031

Rand, David G (2012) 'The Promise of Mechanical Turk: How Online Labor Markets can Help Theorists run Behavioral Experiments.' *Journal of Theoretical Biology* 299, 172–179

Richter, David, Julia Roher, Maria Metzing, Wiebke Nestler, Michael Weinhardt, and Jürgen Schupp (2017) 'SOEP Scales Manual (updated for SOEP-Core v32. 1).' Technical Report, SOEP Survey Papers

Searles, Kathleen, and John Barry Ryan (2015) 'Researchers are Rushing to Amazon's Mechanical Turk. Should They?' *Available at h ttp://www. washingtonpost. com/blogs/monkey-cage/wp/2015/05/04/r esearchers-are-rushing-to-amazons-mechanical-turk-should-they*

Shapiro, Danielle N, Jesse Chandler, and Pam A Mueller (2013) 'Using Mechanical Turk to Study Clinical Populations.' *Clinical Psychological Science* 1(2), 213–220

Stanford, Matthew S, Charles W Mathias, Donald M Dougherty, Sarah L Lake, Nathaniel E Anderson, and Jim H Patton (2009) 'Fifty Years of the Barratt Impulsiveness Scale: An Update and Review.' *Personality and Individual Differences* 47(5), 385–395

Verkoeijen, Peter P. J. L., and Samantha Bouwmeester (2014) 'Does Intuition Cause Cooperation?' *PloS ONE.* https://doi.org/10.1371/journal.pone.0096654

Zaller, John (1992) *The Nature and Origins of Mass Opinion* (New York: Cambridge University Press)