The Micro-Task Market for "Lemons": Collecting Data on Amazon's Mechanical Turk

Douglas J. Ahler* Carolyn E. Roush† Gaurav Sood‡

September 18, 2018

Amazon's Mechanical Turk (MTurk) has rejuvenated the social sciences. It has freed researchers from the lab and from the "narrow database" of social science undergraduates (Sears 1986). It has also dramatically reduced the costs of running a study (e.g., Berinsky, Huber and Lenz 2012; Casler, Bickel and Hackett 2013; Paolacco and Chandler 2014). But the days of cheap, good data may be coming to an end. Recent evidence suggests that a non-trivial proportion of data collected on MTurk is "suspicious," generated either by "non-respondents" (bots) or non-serious respondents (Bai 2018; Ryan 2018).

Spurred by the concern, we decided to probe the issue further. In August 2018, we fielded an original survey on MTurk in which we included questions designed to measure non-serious responses. In addition, we used a Qualtrics plugin to record the IP addresses of the devices from which responses were being filed. We then augmented the data by collecting IP-level metadata, such as the estimated location of the device, from various services using Know Your IP. We then used the IP-metadata to study fraudulent responses.

We find that about 11% of respondents likely circumvented location requirements or used multiple devices from the same IP address. Roughly 16% of the responses came

*Assistant Professor of Political Science, Florida State University, dahler@fsu.edu

†Democracy Postdoctoral Fellow, the Ash Center for Democratic Governance and Innovation at the Harvard Kennedy School, carolyn_roush@hks.harvard.edu

‡Gaurav can be reached at: gsood07@gmail.com

from blacklisted IP addresses. Finally, we estimate that between 5–7% of the respondents responded non-seriously, either because they were satisficing or because they were "trolling." In all, we find that about 25% of the responses we received are potentially untrustworthy.

Incentives For Quality

MTurk is a micro-task market: people work on small `Human Intelligence Tasks` (HITs) for small amounts of money. MTurk maintains ratings on all users, which means that both `Requestors` (employers) and `Workers` have incentives to behave. `Requestors` have incentives to fairly represent the nature of work being offered, pay a competitive wage, pay up promptly, and not withhold payments unjustly; `Workers` have incentives to submit high-quality work.

Incentives for quality, however, vary by how hard it is to observe quality (Akerlof 1978). For example, workers have smaller incentives to be honest about things like demographic information or the location from which they are taking the survey, as `Requestors` often cannot directly observe these characteristics. `Workers` may exploit this opacity for gain. For example, foreign nationals may complete HITs limited to Americans because such HITs tend to be more lucrative, given differences in purchasing power parity. Another possibility is that `Workers` may create multiple accounts and complete the same HIT multiple times even when it is explicitly stated that a person may only complete each HIT once.

But these are just two examples; the problem is more general. On a whole host of tasks, the quality of work is hard to observe. This point holds for surveys. Except for cases where a respondent takes extraordinarily little time to finish a survey, we cannot definitively say that someone is not putting in effort when answering questions. For instance, selecting the first response option of multiple questions in a row is not conclusive evidence of satisficing (Krosnick, Narayan and Smith 1996; Vannette and Krosnick 2014), and respondents may exploit this ambiguity by rushing through the survey. In all, `Workers` may deliberately

misrepresent where are they are located, masquerade as someone else when "double dipping," or complete surveys insincerely and relatively quickly, perhaps even by using software to autofill forms.[1]

Beyond this, the difficulty in measuring the quality of responses also means that there are weak incentives for `Workers` to take surveys seriously. People who enjoy saying outrageous things or misleading interviewers may use the opportunity to make a few bucks while indulging their id (Cornell et al. 2012; Lopez and Hillygus 2018; Robinson-Cimpian 2014; Savin-Williams and Joyner 2014).

Until now, we have talked about incentives for quality within a HIT. But what is the market signal? For `Workers`, it is the percentage of HITs—of all kinds, not just surveys—that are approved. Because of the previously discussed difficulties in judging quality, workers' quality signal is upwardly biased. This may especially be the case because of a market signal for `Requestors`: `Workers` can rate HITs and view `Requestors`' average ratings using Turkopticon. `Requestors`, therefore, face increased incentives to be cautious and generous when approving responses, and workers' quality signal may be especially biased toward the 95% threshold some `Requestors` use as a filter for who may take their surveys. This filter likely keeps out the worst serial offenders, but we strongly suspect that `Workers`' approval rating is a noisy signal of the quality of responses they are likely to provide.

Data and Measures

To study how common these types of problematic responses are on MTurk, we posted a survey on August 17th, 2018, advertising the HIT as "30 short questions on various topics on education, learning, and American society." We solicited 2,000 responses from MTurk `Workers` located in the United States. We paid $0.60 for each completed HIT.

[1] You can find examples of these kinds of programs here or here.

To identify people masquerading as someone else or providing misleading answers about their location, we exploited data on IP addresses. We used a built-in Qualtrics plugin to collect respondents' IP addresses. We then used Know Your IP (Laohaprapanon and Sood 2018), which provides a simple interface to multiple services to obtain metadata on IP addresses. In particular, we got the location—the estimated latitude and longitude, the city and the country—from MaxMind (MaxMind 2006), the largest, most trusted provider of geoIP data. We also used ipvoid.com, which collates data from 96 separate blacklists, to check whether or not an IP address is on a "blacklist." IP addresses are blacklisted for two main reasons: (1) if a website associated with the IP is caught spreading malware or engaging in phishing, or (2) bad Internet traffic like a DDoS attack originates from the IP. While IP addresses are not static—over time, they can be re-used—the turnover rate is generally slow, which means we can be fairly certain that the IP addresses appearing on blacklists have not been recently reassigned to other users. We suspect that responses from blacklisted IPs come from people using VPNs—or some similar web service—to log in to MTurk using U.S. IP addresses (and thus circumvent location filters), but we cannot be sure.

To determine how many workers were bypassing our location requirement, we examined the IP addresses to pinpoint which workers completed the survey from outside the United States. To determine how many people were using potentially using multiple accounts to complete HITs more than once, we also counted the number of responses per IP address.

In addition to this, we used `No CAPTCHA reCAPTCHA` (Shet 2014), which uses mouse movements on the screen to estimate whether they are being made by a human or a bot—that is, to check if respondents are using bots or other auto-fill tools to complete surveys.

To identify satisficers and "trolls," i. e., people who intentionally provide humorous or provocative responses, we followed Lopez and Hillygus (2018) by asking respondents about rare afflictions, behaviors, and traits, as well as an explicit question about how sincerely they

respond to surveys (Cornell et al. 2012; Robinson-Cimpian 2014; Savin-Williams and Joyner 2014). In particular, we asked respondents whether or not they belonged to a gang, had an artificial limb, were blind or had impaired vision, and whether or not they had a hearing impairment. We also asked respondents how much they slept and coded anyone reporting sleeping more than ten hours or fewer than four hours as unusual. (For exact question text, see SI 1.1.) In keeping with previous research, we flag respondents as satisficing or trolling if they answer "yes" or reported unusual behavior on two or more of these questions (Lopez and Hillygus 2018).[2] Lastly, we examine how responses to these questions correlate with self-reports of how honest they generally are in answering surveys.

Results

First, some good news. All respondents who were asked to confirm that they were human using `NoCaptcha ReCaptcha` passed. That is all the good news we have; the rest of the data make for grim reading.

Of the 2,000 responses, the Qualtrics plugin was able to record the IP addresses of 1,991 responses. (We consider the nine responses for which Qualtrics couldn't record the IP address as suspect.) Of the 1,991 responses, 106 responses were submitted from an IP that appears in our dataset more than once (see Table SI 1.1). This could be because multiple people in the same household filled out the HIT—this information would still be useful for survey researchers as it affects standard error calculations—but we think the more plausible explanation is that respondents used multiple accounts to submit the same HIT multiple times.

1,866 responses originated from within the United States (see Table 1). Of the 125 re-

[2]It is plausible, even likely, that people with physical disabilities or those that come from marginalized groups are overrepresented on MTurk. Ideally, we would like to have more defensible priors about the true proportion of say, blind gang members in our sample. Without it, we cannot be certain.

sponses filed from outside the United States, 42 were from Venezuela and 17 were from India. (See Table SI 1.2 for a fulll distribution of countries from which the HIT was completed.) We suspect that these 125 responses are from MTurk `Worker` accounts that were created using U.S. credit cards but belong to people living in other countries. It is plausible that the foreign IP addresses represent Americans who are currently traveling but the geographic distribution of the IP addresses suggests this is unlikely.

In looking at data at the city level, another puzzling fact emerges. Consistent with Ryan (2018), the city from which the most responses were filed is Buffalo, with 77 responses. (Table SI 1.3 shows all the cities from which more than 10 responses were filed.) The other cities are either big American cities or a city in Venezuela. Geolocation at the city level is not reliable enough to definitively say that this pattern is problematic but there are good reasons to be suspicious.

Yet more shockingly, of the 1,991 responses, 321 come from blacklisted IPs. In all, 408 responses—or around 20% of the sample—came from outside the United States, blacklisted IP addresses, duplicate IPs, or missing IPs.

Type of Suspicious IP	n
Missing	9
Blacklisted	321
Duplicated	106
Foreign	125
Any of the Above	408

Table 1: Frequency of Different Types of Suspicious IPs

Next, we examined the frequency of insincere responses (which, alternatively, may indicate that the respondent was not paying attention). Just over 9% of respondents report being blind or having a visual impairment (see Table 2). Another 5.5% report being deaf. These numbers are nearly three times and 14.5 times the respective rates in the population.[3]

[3]Less than half of a percent of Americans aged five or older are deaf (Mitchell 2005) and about 3% of Americans 40 or older are blind or visually impaired (CDC).

These large deviations from the national norm are possible but unlikely. Questions on gang membership have similarly implausible numbers. About 6% respondents report having a family member in a gang. This compares to a rate of about half a percent in the overall population (National Gang Intelligence Center (U.S.) 2012). These are clearly implausible numbers. But to be cautious, we only flag a respondent as potentially engaging in trolling if she provided a "yes" response on two or more on such items. (See Figure 1 for the distribution of affirmative responses to these questions.) In all, there are a total of 125 such responses (or a little over 6% of all responses). Additionally, 99 respondents (or roughly 5% of the sample) reported that they "always" or "almost always" provided humorous or insincere responses to survey questions.

Rare Behaviors/Traits	n
Use a Prosthetic	91
Blind	184
Deaf	109
Gang Member	88
Family Member in Gang	123
Sleep 10+ hrs or <4 hrs	28
2 or more of above	125

Table 2: Number of People Who Report Having Rare Behaviors/Traits

When we look at the association between admittedly providing insincere responses and responding affirmatively to questions about rare afflictions, traits, and behaviors, an expected pattern emerges. Among the 1,875 respondents who marked yes on one or fewer on questions about rare things, about 93% said that they "never" or "rarely" provided humorous or insincere responses. By contrast, among the 125 who scored two or higher on these questions, about 58% said they were usually sincere in their responses ($\chi^2 = 179.0, p < .001$). In all, we suspect that 5–7% of the `Workers` recruited for this study engaged in trolling or satisficing.

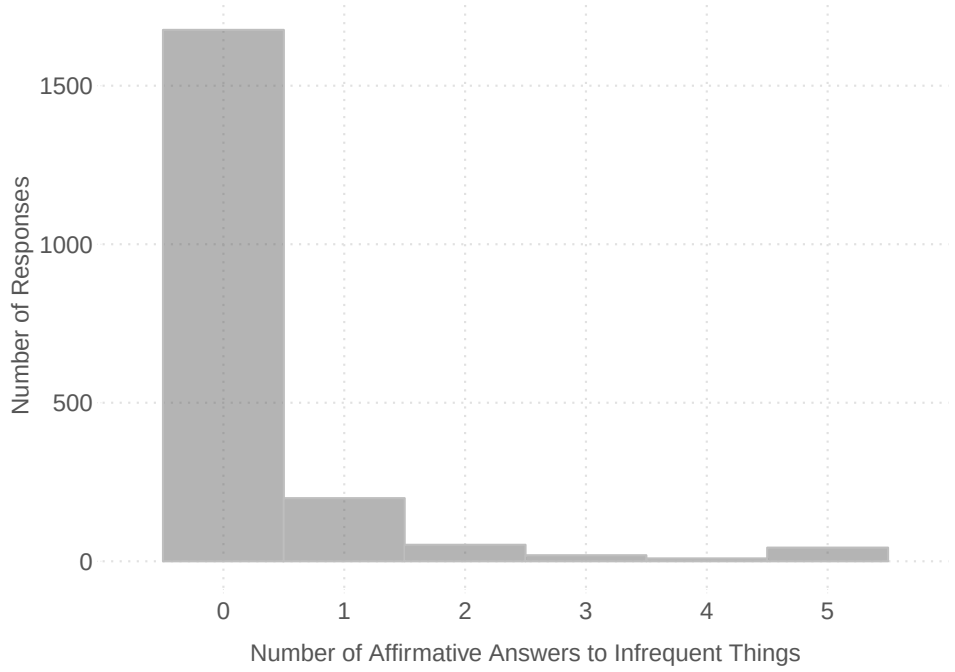

Figure 1: Distribution of Affirmative Responses to Low-Incidence Screener Questions

Finally, we cross-referenced the data on IP addresses with responses to questions about rare behaviors, traits, and conditions. Thirty eight (38) of the 408 responses (or about 9% of the responses) from "bad" IPs also replied in the affirmative on two or more of such questions, while nearly 6% of the responses from other IPs did the same. The difference is statistically significant but not eye-catchingly large. We didn't necessarily expect it to be: people who earn money on MTurk want to do enough to get paid and not get barred by Amazon. The question is if we want data from these actors. We think not.

In all, 495 responses are from IPs that are duplicated, located in a foreign country, or blacklisted, or provide affirmative answers to two or more of the low-incidence questions. In sum, nearly a quarter of the responses we recorded are potentially untrustworthy.

Discussion

The results suggest that there are three significant concerns with the quality of survey data being collected on Amazon's Mechanical Turk. First, a non-trivial proportion of responses are from people who are not in the sampling frame. Second, some people are likely filing multiple responses. Third, a significant proportion of people are responding non-seriously.

We cannot estimate the consequence of the first problem because the structure of error added to the data-generating process is unknown. When we receive multiple responses from the same IP address, at the bare minimum, we obtain misleading estimates of standard errors. For the third issue, we expect at least two consequences. To the extent that people are answering "randomly," the consequences are: a) attenuated correlations, and b) biased estimates of frequencies and means on some variables. For instance, even random answering can positively bias estimates of how many people know something (Cor and Sood 2016). If people are answering humorously or with the aim of being provocative, they only contribute to bias.

Current data quality may be low, but what's the prognosis? Given the strategic incentives, we forecast steadily declining quality. Unless we can craft and implement better methods to assess and incentivize quality responding, the chances that things will improve seem low. Ultimately, it is important that the methods we devise preclude new ways of gaming the system, or we are back to square one. For now, we can think of only a few recommendations for researchers:

- Use geolocation filters on survey platforms like Qualtrics to enforce any geographic restrictions.
- Make use of tools on survey platforms to retrieve IP addresses. Run each IP through Know Your IP to identify blacklisted IPs and multiple responses originating from the same IP.

- Include questions to detecting trolling and satisficing, but don't copy and paste from a standard canon as that makes "gaming the survey" easier.

- Caveat emptor: increase the time between HIT completion and auto-approval so that you can assess your data for untrustworthy responses before approving or rejecting the HIT. We approved all HITs here because we used all the responses in the analysis. But researchers may decide to only pay for responses that pass some low bar of quality control. But caveat lector: any quality control must pass two tough tests: 1) it should be fair to the `Workers`, and 2) it should not be easily gamed.

  Rather than withhold payments, a better policy may be to incentivize workers by giving them a bonus when their responses pass quality filters. This would lead to a weak signal propagating in the market, where people who do higher quality work are paid more, and eventually come to dominate the market. If multiple researchers agree to provide such incentives around reliable quality checks immune to being gamed, we may be able to change the market. Another possibility is to create an alternate set of ratings for `Workers` not based on HIT approval rate—much like how `Workers` use Turkopticon to assess `Requestors`' generosity, fairness, etc.

- Use worker qualifications on MTurk to filter only `Workers` who have a high percentage of approved HITs into your sample. Since we expect sharp upward bias in the metric, we think filtering on upper-90s may be par for the course. Over time, this may also change the market.

Lastly, we would like to note that we do not think that the problem is limited to MTurk. MTurk may be more prone to "lemon" responses because: a) it is a market with multiple independent employers rather than one central respondent management system, and b) the only signal of response quality that is propagated to the market is HIT approval. But

on any paid platform, including those that offer small incentives, we think it is a potential concern.

References

Akerlof, George A. 1978. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. In *Uncertainty in Economics*. Elsevier pp. 235–251.

Bai, Hui. 2018. "Evidence that a Large Amount of Low Quality Responses on MTurk Can be Detected with Repeated GPS Coordinates." https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random.

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.

Casler, Krista, Lydia Bickel and Elizabeth Hackett. 2013. "Separate but Equal? A Comparison of Participants and Data Gathered via Amazon's MTurk, Social Media, and Face-to-Face Behavioral Testing." *Computers in Human Behavior* 29(6):2156–2160.

Cor, M Ken and Gaurav Sood. 2016. "Guessing and forgetting: A latent class model for measuring learning." *Political Analysis* 24(2):226–242.

Cornell, Dewey, Jennifer Klein, Tim Konold and Frances Huang. 2012. "Effects of Validity Screening Items on Adolescent Survey Data." *Psychological Assessment* 24(1):21–35.

Krosnick, Jon A., Sowmya Narayan and Wendy R. Smith. 1996. "Satisficing in Surveys: Initial Evidence." *New Directions for Evaluation* 70:29–44.

Laohaprapanon, Suriyan and Gaurav Sood. 2018. "Know Your IP.".
URL: *https://github.com/themains/know_your_ip*

Lopez, Jesse and D. Sunshine Hillygus. 2018. Why So Serious? Survey Trolls and Misinformation. In *Annual Meeting of the Midwest Political Science Association*.

MaxMind, LLC. 2006. "GeoIP.".

Mitchell, Ross E. 2005. "How Many Deaf People are There in the United States? Estimates from the Survey of Income and Program Participation." *Journal of Deaf Studies and Deaf Education* 11(1):112–119.

National Gang Intelligence Center (U.S.). 2012. *2011 National Gang Threat Assessment: Emerging Trends*. New York, NY.

Paolacco, Gabriele and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23(3):184–188.

Robinson-Cimpian, Joseph P. 2014. "Inaccurate Estimation of Disparities Due to Mischevious Responders: Several Suggestions to Assess Conclusions." *Educational Researcher* 43(4):171–185.

Ryan, Timothy J. 2018. "Data Contamination on MTurk." http://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/.

Savin-Williams, Ritch C. and Kara Joyner. 2014. "The Dubious Assessment of Gay, Lesbian, and Bisexual Adolescents of Add Health." *Archives of Sexual Behavior* 43(3):413–422.

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51(3):515–530.

Shet, Vinay. 2014. "Are You a Robot? Introducing 'No CAPTCHA reCAPTCHA'." *Google Security Blog* 3:12.

Vannette, David L. and Jon A. Krosnick. 2014. A Comparison of Survey Satificing and Mindlessness. In *The Wiley Blackwell Handbook of Mindfulness*, ed. Amanda Ie, Christelle T. Ngnoumen and Ellen J. Langer. Malden: Wiley pp. 312–327.

SI 1 Supporting Information

Freq	Count
1	1885
2	20
3	13
4	4
5	1
6	1

Table SI 1.1: Number of times an IP Address is in the data

Country	Freq
United States	1866
Venezuela	42
India	17
Canada	6
Puerto Rico	4
Brazil	3
Honduras	3
Kenya	3
Philippines	3
Albania	2
Ecuador	2
Egypt	2
Germany	2
Mexico	2
Nepal	2
Tajikistan	2
Thailand	2
United Kingdom	2
Uzbekistan	2
Vietnam	2
Argentina	1
Chile	1
Colombia	1
Czechia	1
Georgia	1
Ghana	1
Greece	1
Guinea	1
Jamaica	1
Macedonia	1
Nigeria	1
Pakistan	1
Portugal	1
Republic of Korea	1
Russia	1
Saint Vincent and the Grenadines	1
Seychelles	1
Suriname	1
Taiwan	1
United Arab Emirates	1

Table SI 1.2: Countries from which responses were recorded.

City	Freq
Buffalo	77
New York	72
Los Angeles	44
Maracaibo	31
Kansas City	28
San Francisco	21
Houston	19
Chicago	18
Brooklyn	17
Miami	16
Charlotte	15
Orlando	15
Columbus	14
Austin	13
Jacksonville	13
Philadelphia	12
Portland	12

Table SI 1.3: Cities from which more than 10 responses were recorded.

SI 1.1 Question Text

- Do you use an artificial limb or prosthetic?—Yes, No
- Are you blind or do you have vision impairment?—Yes, No
- Are you deaf or do you have hearing impairment?—Yes, No
- Are you in a gang?—Yes, No
- Is one or more of your immediate family members in a gang?—Yes, No
- Finally, we sometimes find people don't always take surveys seriously, instead providing humorous, or insincere responses to questions. How often do you do this? — Never, Rarely, Some of the time, Most of the time, Always