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

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

November 17, 2018

Abstract

Amazon's Mechanical Turk (MTurk) has rejuvenated the social sciences, dramatically reducing the costs of collecting original data. Recent evidence, however, suggests that a non-trivial portion of data collected on MTurk is "suspicious," either generated by "non-respondents" (bots) or non-serious respondents. Spurred by this concern, we fielded an original survey in August 2018 designed to measure the prevalence of both "cheating" and "trolling" on the platform. We find that about 20% of respondents likely circumvented location requirements, used multiple devices to fill out the same survey, or completed the HIT from a blacklisted IP address. In addition, we estimate that about 5-7% of our sample provided non-serious responses to our questions, either because they were satisficing or trolling. Altogether, we find about a quarter of our responses to be potentially untrustworthy. Perhaps more troublingly, we also find evidence that cheating and trolling moderates treatment effects by introducing noise into dependent measures, a form of attenuation bias. We conclude by providing some recommendations for researchers to improve data collection and quality on MTurk.

*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

Amazon's Mechanical Turk (MTurk) has rejuvenated the social sciences. It has freed researchers from the a reliance on the "narrow database" of social science undergraduates (Sears 1986) while dramatically reducing the cost of conducting a study (e.g., Berinsky, Huber and Lenz 2012; Casler, Bickel and Hackett 2013; Paolacco and Chandler 2014). For these reasons, the platform has become a particularly popular venue for social science experimentation. Though MTurk's convenience samples may not generalize to a broader population, random assignment reduces selection bias by definition (Gerber and Green 2012; Shadish, Cook and Campbell 2002), making the platform a cheap and easy way for researchers to test causal hypotheses. Moreover, on the surface, other threats to internal validity appear minimal. MTurk participants—`Workers`—appear to be as attentive as subjects in lab studies (e.g., Hauser and Schwarz 2016; Mullinix et al. 2015; Thomas and Clifford 2015) and exhibit the same cognitive biases as subjects recruited through more traditional means (e.g., Goodman, Cryer and Cheema 2012; Horton, Rand and Zeckhauser 2011; Paolacci, Chandler and Ipeirotis 2010). Consequently, MTurk data tends to produce treatment effects that approximate those produced in other convenience and nationally-represented samples (e.g., Mullinix et al. 2015; Thomas and Clifford 2015). MTurk has therefore been seen as an inexpensive means through which researchers can gather high-quality experimental data to aid them in making scientific inferences about human behavior.

But the days of cheap, good data may be coming to an end. Recently, several researchers have discovered that a non-trivial proportion of data collected on MTurk is "suspicious," generated either by "non-respondents" (bots) or non-serious respondents (e.g., Bai 2018; Dreyfuss 2018; Ryan 2018). While this type of fraudulent behavior has attracted attention only recently, we believe it is symptomatic of a larger problem on MTurk. MTurk's market structure incentivizes `Workers` to complete as many HITs as possible to maximize their payout, and social scientists—who primarily use the platform to solicit opinions, not gauge objective performance—have little means by which to determine whether `Workers` are

misrepresenting themselves or their attitudes. This means that existing signals of quality that researchers rely on to parse "good" `Workers` from "bad"—i.e. the HIT completion rate—are artifically inflated, as they do not and can not account for common types of insincere responding like cheating, satisficing, or trolling.

These incentive structures mean that the presence of low-quality respondents on MTurk is potentially far more widespread than anticipated. This is particularly problematic for experimentalists, as an epidemic of fraudulent responding on MTurk may change the substantive conclusions reached by an experiment. For example, if bots or survey satisficers provide more or less random answers to questions, they could bias average treatment effects toward zero. On the other hand, if respondents intentionally select the most humorous or outrageous response to a question—that is, they engage in "trolling"—they could systematically distort estimates in other ways, such as upwardly biasing estimates of particular contours of opinion, like belief in misinformation (e.g. Lopez and Hillygus 2018). In either case, the presence of a large number of insincere or fraudulent respondents on the platform calls into question the utility of MTurk as a potential venue for experimental research.

Spurred by these concerns, we decided to probe the issue further. In August 2018, we fielded an original study on MTurk in which we gauged the prevalence of fraudulent responding and its potential impact on experimental results. To identify respondents masquerading as someone else, 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, to more closely examine suspicious responses. In addition, we included a battery of questions designed to measure satisficing or trolling in an effort to determine how many `Workers` responded non-seriously to questions on our survey.

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

blacklisted IP addresses. In addition, we estimate that between 5–7% of our respondents engaged in satisficing or trolling. In all, we find that about 25% of the responses we received are potentially untrustworthy.

We also provide evidence that these suspicious responses bias experimental results. Using a vignette experiment embedded in the same survey, we find that respondents who cheat or troll differ from ordinary survey-takers in how they respond to treatments. Specifically, suspicious responding attenuates treatment effects by introducing significant noise into the data. This suggests that low-quality data may be responsible for Type II errors in studies conducted using data from MTurk.

While cheating, trolling. and/or satisficing appear to be widespread on MTurk, we are nevertheless optimistic that the platform can continue to be used to conduct scientific research. While current data quality may be low, we believe researchers can craft and implement better methods to assess and incentivize quality responses. We conclude with a few recommendations for researchers about how to minimize susppicious—and potentially fraudulent—responding in their studies.

Incentives For Quality on MTurk

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` (participants) 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). `Workers` have relatively few 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, work quality is difficult to observe. This is particularly true for surveys. While `Requestors` can gauge `Workers'` objective *performance* on some tasks—for example, by spot checking coding decisions for agreement with others (e.g., Garz et al. 2018)—most social scientists use MTurk to solicit `Worker` *opinions*, which by definition have no "correct" answer. This makes it incredibly difficult to parse genuine responses from fraudulent ones. For example, except for cases where a respondent takes extraordinarily little time to finish a survey, reseachers cannot definitively say gauge whether the respondent is or is not putting in effort when answering questions. Even 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). Respondents could exploit this ambiguity by rushing through the survey in order to complete it—and thereby receive their reward—as soon as possible.

`Workers`, therefore, have strong incentives to 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] The ambiguity in identifying quality responses also means MTurk may be particularly attractive to people who enjoy providing outrageous or misleading responses within surveys or elsewhere on the internet (i.e. "trolls"), as it allows them to make money while indulging their id

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

(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% completion threshold some `Requestors` use as a filter for who may take their surveys (e.g., Peer, Vosgerau and Acquisti 2014). 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.

All of this suggests that data collected on MTurk may not be of as high quality as researchers may assume. There are distinct incentives for `Workers` to misrepresent themselves and their beliefs, and the degree to which `Workers` engage in this type of behavior may not be captured by existing quality control protocols or reflected in signals of `Worker` quality. Taken together, this suggests that the presence of low-quality responding on MTurk may be far more prevalent than is commonly understood.

Assessing the Prevalence of Suspicious Behavior

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. In keeping

with best practices, we restricted participation to MTurk workers with a high reputation (that is, a HIT completion rate of at least 95%) (Peer, Vosgerau and Acquisti 2014).

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 that draws from multiple services to obtain metadata on IP addresses. In particular, Know Your IP retrieved the location of the `Workers'` IP addresses—the estimated latitude and longitude, the city and the country—from MaxMind (MaxMind 2006), the largest, most trusted provider of geoIP data. It also uses 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. To assess the prevalence of bots on the platform, 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 computer—that is, to check if respondents are using bots or other auto-fill tools to complete surveys.

To identify satisficers and "trolls," we followed Lopez and Hillygus (2018) in 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 (and whether or not an immediate family member did), 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 respondents generally are in answering surveys.

First, some good news. All respondents who were asked to confirm that they were human using `NoCaptcha ReCaptcha` passed. This suggests that concerns about a "bot panic" (Dreyfuss 2018) on MTurk may be overwrought. However, this 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 completed 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.

[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.

A majority of responses (1,866) originated from within the United States (see Table 1). Of the 125 responses 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.

Table 1: Frequency of Different Types of Suspicious IPs

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

In looking at data at the city level, another puzzle 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, consistent with findings from Kennedy et al. (2018). 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.

Next, we examined the frequency of insincere or potentially inattentive respondents. Just over 9% of respondents in our data 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] 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. To be cautious, however, 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.

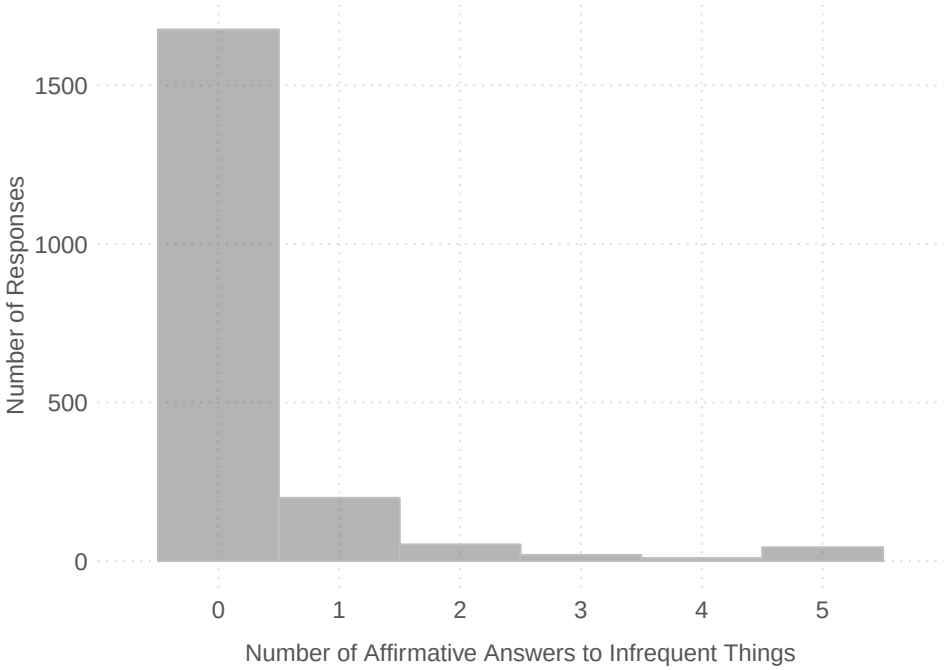
Table 2: Number of People Who Report Having Rare Behaviors/Traits

Rare Behavior/Trait	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

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.

[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).

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



Finally, we cross-referenced the data on IP addresses with the responses to questions about rare behaviors, traits, and conditions. 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. In comparison, nearly 6% of the responses from other, non-suspicious IPs did the same. The difference is statistically significant but not eye-catchingly large. We didn't necessarily expect it to be, as people who earn money on MTurk want to do enough to get paid and not get barred by Amazon. Whether we want data from these actors, however, is another question.

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.

Consequences of Suspicious Behavior

The results above suggest that there are three significant concerns with the quality of survey data being collected on MTurk. 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. The second issue is more problematic. When we receive multiple responses from the same IP address, at the bare minimum, we obtain misleading estimates of standard errors.

The consequences of other suspicious behavior are potentially more grave. To the extent that `Workers` provide more or less "random" responses to questions, they can introduce noise into the data that results in attenuated correlations, which can bias the 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, on the other hand, people are answering humorously or with the aim of being provocative, they may introduce more *systematic* bias into estimates of the prevalence of certain attitudes. Depending on the relationship between this bias and variables of interest, estimates could either be overinflated or underinflated. Both circumstances pose serious threats to researchers' ability to draw accuate scientific inferences.

To examine the extent to which suspicious responding influences results, we investigated whether suspicious and non-suspicious respondents react differently to experimental stimuli. To do so, we embedded an experiment on partisan stereotyping—one with predictable treatment effects—into the aforementioned survey.

We borrow a design from Ahler and Sood (2017), itself a modification of the famous "Linda Problem" (Tversky and Kahneman 1974). In the study, Ahler and Sood (2017) ex-

amine how individuals' reliance on the representativeness heuristic—that is, the tendency to associate distinguishing traits with groups, regardless of other information available—drives Americans' perceptual bias about how the Democratic and Republican parties' bases are composed. Specifically, they examine how exposure to party-representative characteristics (like gender, race, sexual orientation, and religion) influences the likelihood that respondents will commit the *conjunction fallacy*, a cognitive error that occurs when people assert that the probability of two events together is greater than the probability of either event occuring separately (Tversky and Kahneman 1974).

To test this hypothesis, Ahler and Sood (2017) presented different versions of a character named James to respondents, randomly and independently manipulating particular characteristics within a vignette. This design is ideal for our purposes here, as the independent manipulation of several features allows for multiple tests of attenuation bias within one experiment—that is, instead of comparing how suspicious and non-suspicious respondents differ in their response to one treatment, we can do so for multiple treatments at once, improving statistical power. The vignette read as follows:

> James is a 37-year-old (`white` | `black`) man. He attended the University of Michigan, where he double-majored in economics and political science. While there, James was president of a business and marketing club. He also participated in (`anti-tax demonstrations` | `living-wage demonstrations` | `student government`).
>
> James's co-workers describe him as highly driven, outspoken, and confident. He is married to (`Karen` | `Keith`) and has one son. In James's free time, he (`leads his son's Cub Scouts group, organized through the Baptist Church the family attends` | `leads his son's Junior Explorers group, led through the Secular Families Foundation` | `coaches his son's youth sports teams`).

Following exposure to the vignette, Ahler and Sood asked respondents what they

believe to be most likely among three options: (1) "James is a salesman," (2) "James is a salesman who also supports the Democratic Party," and (3) "James is a salesman who also supports the Republican Party." Of course, the latter two options are logically impossible, as the probability that James is both a salesman and a supporter of the Republican Party will always be less than or equal to the probability that he is either a salesman or a member of the Republican Party. Thus, in selecting option (2) or (3), respondents commit the conjunction fallacy. Ahler and Sood (2017) find, unsurprisingly, that characteristics that are representative (Tversky and Kahneman 1974) of the Democratic (Republican) Party lead individuals to commit the Democratic (Republican) conjunction fallacy.

To test whether suspicious respondents differ in their response to the treatments, we estimated the *average marginal component effect* (AMCE) of each independently randomized characteristic, interacted with an indicator for suspicious behavior, on the probability that respondents make the Democratic and Republican conjunction fallacies. Since the dependent variable takes on three values—Democratic conjunction fallacy (-1), logically correct response (0), Republican conjunction fallacy (1)—we use an ordered logit model to analyze the data.[4] Thus, our model takes the following form:

$$p_{ij} = p(y_i = j) = \begin{cases} p(y_i = -1) = p(y_i^* \leq \alpha_{-1}) \\ p(y_i = 0) = p(\alpha_{-1} < y_i^* \leq \alpha_0) \\ p(y_i = 1) = p(\alpha_0 < y_i^*) \end{cases} \tag{1}$$

where y_i^* is the respondent's latent outcome and α_{-1} and α_0 are the model's cutpoints. We model these probabilities as follows:

$$p(y_i = j) \sim \text{logit}^{-1}(\beta_k X_{ik} + \delta S_i + \gamma(S_i \times X_{ik}) + \varepsilon) \tag{2}$$

[4]We omit one value per variable in this model.

where X_k denotes our vector of randomly and independently assigned characteristics of James and S_i is an indicator for "suspicious respondent."[5]

Full model results are available in SI 1.2. For ease of interpretation, we present marginal effects in Table 3, specified as the change in predicted probability of committing the Democratic/Republican conjunction fallacy. We first present results for all non-suspicious respondents (column 1) and then by all suspicious respondents (which include flagged IP addresses and respondents we suspect are trolling) (column 2). Finally, we present the results among flagged IP addresses alone (column 3) and non-serious respondents alone (column 4).

The first column confirms significant average marginal component effects (AMCEs) of all randomly and independently varied characteristics. Non-suspicious respondents are significantly more likely to commit the Democratic conjunction fallacy when James is presented as black, gay, secular, or described as having liberal policy preferences; they are also more likely to commit the Republican conjunction fallacy when James is presented as evangelical or described as having conservative policy preferences. In sum, people appear to stereotype others as partisan on the basis of social and policy cues, even making illogical inferences in the process.

Column 2 demonstrates that suspicious respondents respond differently to the treatments. Specifically, AMCEs are generally attenuated among respondents flagged for any reason. The magnitude of the difference between non-suspicious and suspicious respondents is notable. Suspicious respondents, for example, are nearly eight percentage points less likely than non-suspicious respondents to make the Democratic conjunction fallacy when James is presented as black; they are almost ten percentage points less likely to make the Democratic conjunction fallacy when James is presented as gay. Similarly, suspicious respondents are less likely than non-suspicious respondents to make the Republican conjunction fallacy

[5]We operationalize suspicious responding three ways in three different models: first as all respondents flagged for any reason, then for duplicated/flagged IP addresses, and finally for respondents flagged for non-serious responses.

Table 3: Impact of Suspicious Responding on Treatment Effects - Marginal Effects

	Non-suspicious respondents ($n = 1446$)		All suspicious respondents ($n = 487$)		Flagged IPs only ($n = 367$)		Non-serious respondents only ($n = 120$)	
When James is described as...	More likely to make Dem. CF by	More likely to make Rep. CF by	More likely to make Dem. CF by	More likely to make Rep. CF by	More likely to make Dem. CF by	More likely to make Rep. CF by	More likely to make Dem. CF by	More likely to make Rep. CF by
Black (vs. white)	**13.6%**	**-9.4%**	*5.8%*	*-4.2%*	**8.8%**	**-6.2%**	*-3.9%*	*3.3%*
Gay (vs. straight)	**18.5%**	**-12.7%**	***8.6%***	***-6.1%***	**11.8%**	**-8.3%**	*-3.5%*	*3.0%*
Evangelical (vs. nothing)	**-5.7%**	**4.1%**	0.8%	-0.5%	-14%	1.0%	11.9%	-10.1%
Secular (vs. nothing)	**6.9%**	**-4.7%**	6.4%	-4.5%	4.0%	-2.8%	18.7%	-15.1%
Liberal (vs. nothing)	**10.2%**	**-6.8%**	*-0.9%*	*0.7%*	*2.4%*	*1.7%*	*-12.4%*	*11.6%*
Conservative (vs. nothing)	**-7.7%**	**5.5%**	**-15.3%**	**11.7%**	**-16.7%**	**12.7%**	-13.1%	11.7%

Estimates in **bold** are significantly different from zero ($p < 0.1$).

Estimates in *italics* are significantly different from those in the non-suspicious respondents column ($p < 0.1$).

when James is presented as either black or gay. Oddly, the effect of the conservative cue is substantively larger among suspicious respondents, but this difference from non-suspicious respondents is not precisely estimated.

Parsing flagged IPs (column 3) from non-serious respondents (column 4), we notice a few interesting patterns. Estimates are generally attenuated among responses with flagged IPs, but among potential non-serious respondents (i.e. trolls or satisficers), we find some puzzling results. For example, non-serious respondents were significantly more likely to profess James to be a *Democratic* salesman when James was described as evangelical, and they were more likely to commit the Republican conjunction fallacy when James was said to have taken a liberal policy position in college. Oddly, however, the effects of the secular and conservative cues were substantively large—larger than those among non-suspicious respondents—and in the correct direction, albeit imprecisely estimated because of the small number of non-serious respondents. While the results among non-serious respondents appear to mostly add noise to our data, these respondents may pose a larger problem if they respond more systematically to other treatments in a way that differs from non-suspicious respondents.

Discussion and Conclusion

The results above suggest that cheating and trolling/satisficing are a significant problem on MTurk. We find that about a quarter of our data is potentially untrustworthy, and that "problematic" respondents on the platform respond differently to experimental treatments than more "honest" respondents. More specifically, suspicious behavior—either in the form of cheating or trolling/satisficing—adds noise to the data, which can result in attenuated treatment effects.

Current data quality may be low, but what's the prognosis? Given 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 do not 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` can use Turkopticon to assess `Requestors'` generosity, fairness, etc.

- Use `Worker` qualifications on MTurk and filter to include 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: (1) it is a market with multiple independent employers rather than one central respondent management system, and (2) 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.

The Belmont Report forever changed social science by clarifying researchers' relationship with study participants, emphasizing that we must treat those who generate our data with respect, beneficence, and fairness. It was a necessary response in a time of reckoning with traumatic treatments and exploitative recruitment practices. We believe that we are currently reckoning with a new problem in our relationship with research participants, a problem that demands we add "respect for data" to the framework that guides this relationship. We do not believe that respect for data is inconsistent with respect for persons, beneficence, and justice. Following the guidelines laid out above—and being clear about expectations of respondents when obtaining their consent—we believe that researchers can include good-faith participants while fairly screening out those who contribute to the data quality problem.

References

Ahler, Douglas J. and Gaurav Sood. 2017. Typecast: Cognitive Roots of Party Stereotyping. In *Annual Meeting of the Midwest Political Science Association*. Chicago: .

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.

Dreyfuss, Emily. 2018. "A Bot Panic Hits Amazon's Mechanical Turk." *Wired* 17 August. https://www.wired.com/story/amazon-mechanical-turk-bot-panic/.

Garz, Marcel, Gaurav Sood, Daniel F. Stone and Justin Wallace. 2018. "What Drives Demand for Media Slant?". Unpublished manuscript, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3009791.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton & Company.

Goodman, Joseph K., Cynthia E. Cryer and Amar Cheema. 2012. "Data Collection in a Flat World: The Strengthsand Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26(3):213–224.

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.

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

Kennedy, Ryan, Scott Clifford, Tyler Burleigh, Philip Waggoner and Ryan Jewell. 2018. "How Venezuela's Economic Crisis is Undermining Social Science Research—About Everything." *Monkey Cage Blog* 7 November. https://www.washingtonpost.com/news/monkey-cage/wp/2018/11/07/how-the-venezuelan-economic-crisis-is-undermining-social-science-research-about-every?utm_term=.8945c0926825.

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.

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.

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

Paolacci, Gabriele, Jesse Chandler and Panagiotis G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5):411–419.

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.

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.

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.

Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2002. *Experimental and Quasi-Experimnental Designs for Generalized Causal Inference.* New York: Houghton Mifflin Company.

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

Thomas, Kyle A. and Scott Clifford. 2015. "The Generalizability of Survey Experiments." *Computers in Human Behavior* 77:184–197.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185:1124–1131.

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

Table SI 1.1: Number of Times an IP Address Appears in the Data

Freq	n
1	1,885
2	20
3	13
4	4
5	1
6	1

Table SI 1.2: Country of Origin for Responses

Country	*n*
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 1.3: Cities With More Than 10 Responses

City	*n*
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

SI 1.1 Question Text - Non-Serious Responding Battery

- 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

SI 1.2 Results of Fully Specified Ordered Logit Model

Table SI 1.4: Impact of Suspicious Responding on Treatment Effects - Full Ordered Logit

	All respondents	Suspicious IPs	Non-serious respondents
Suspicious response	-0.32	-0.36	-0.24
	(0.26)	(0.30)	(0.59)
Black	-0.60	-0.59	-0.59
	(0.10)	(0.10)	(0.10)
Black * SR	0.36	0.24	0.78
	(0.21)	(0.24)	(0.43)
Gay	-0.80	-0.80	-0.80
	(0.10)	(0.10)	(0.10)
Gay * SR	0.46	0.32	0.94
	(0.21)	(0.23)	(0.43)
Evangelical	0.25	0.25	0.25
	(0.12)	(0.12)	(0.12)
Evang. * SR	-0.28	-0.19	-0.75
	(0.25)	(0.28)	(0.56)
Atheist/agnostic	-0.30	-0.30	-0.30
	(0.13)	(0.13)	(0.13)
AA * SR	0.04	0.14	-0.46
	(0.25)	(0.29)	(0.53)
Liberal	-0.45	-0.45	-0.45
	(0.13)	(0.13)	(0.13)
Lib. * SR	0.49	0.54	0.98
	(0.25)	(0.29)	(0.53)
Conservative	0.34	0.34	0.33
	(0.12)	(0.12)	(0.12)
Con. * SR	0.28	0.35	0.22
	(0.25)	(0.29)	(0.52)
Cut 1	-0.60	-0.59	-0.59
	(0.13)	(0.13)	(0.13)
Cut 2	0.66	0.65	0.65
	(0.14)	(0.14)	(0.14)
Pseudo R^2	0.04	0.04	0.05
n	1933	1813	1566

NOTE: "SR" is an indicator for "suspicious respondent.." Its exact operationaliztion changes from model to model. In Column 1, SR == 1 includes all respondents flagged for any reason. In Column 2 we drop likely non-serious respondents so that SR == 1 only includes respondents flagged for suspicious IP addresses. Finally, in Column 3 we drop respondents flagged for suspicious IP addresses so that SR == 1 only includes respondents flagged as potential trolls.