The Micro-Task Market for Lemons: Data Quality on Amazon's Mechanical Turk*

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

May 20, 2019

Abstract

Amazon's Mechanical Turk (MTurk) has rejuvenated the social sciences, dramatically reducing the cost and inconvenience of collecting original data. Recently, however, researchers have raised concerns about the presence of "non-respondents" (bots) and non-serious respondents on the platform. Spurred by these concerns, we fielded an original survey on MTurk to measure response quality. While we find no evidence of a "bot epidemic," we do find that a significant portion of survey respondents behaved suspiciously while taking the survey. About 20% of respondents either circumvented location requirements or took the survey multiple times, and at least 5-7% of participants likely engaged in "trolling" or satisficing. Altogether, we find about a quarter of data collected on MTurk is potentially untrustworthy. Furthermore, we find evidence suggesting that the prevalence of low quality responses in MTurk data has increased over time, and that low quality responses are more common on MTurk than on other online survey platforms. Finally, we find response quality impacts experimental treatments. On average, low quality responses attenuate treatment effects by approximately 10%. We conclude by providing recommendations for collecting data on MTurk.

*This title is inspired by George Akerlof's (1970) seminal paper on quality uncertainty in a market, "The Market for Lemons." We are grateful to Alexander Adams, Don Green, Stephen Goggin, Jonathan Nagler, and John Sides for the useful comments and suggestions. A previous version of this paper was presented at the 12th Annual NYU-CESS Experimental Political Science Conference, February 8, 2019 and at the Annual Meeting of the Midwest Political Science Association, April 6, 2019.

†Corresponding author, 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

1 Introduction

Over the past decade, Amazon's Mechanical Turk (MTurk) has revolutionized experimental social science. The platform has freed researchers from reliance on the "narrow database" of social science undergraduates (Sears 1986) while dramatically reducing the cost and inconvenience of testing causal hypotheses (e.g., Berinsky, Huber and Lenz 2012; Casler, Bickel and Hackett 2013; Paolacco and Chandler 2014). While respondents recruited on MTurk are not representative of the broader population, they are about as attentive as lab subjects (e.g., Hauser and Schwarz 2016; Mullinix et al. 2015; Thomas and Clifford 2015) and exhibit the same cognitive biases as participants recruited through more traditional means (e.g., Goodman, Cryer and Cheema 2012; Horton, Rand and Zeckhauser 2011; Paolacci, Chandler and Ipeirotis 2010). It is perhaps unsurprising, then, that treatment effects on MTurk tend to approximate those found in other convenience and population-representative samples (e.g., Mullinix et al. 2015; Thomas and Clifford 2015).

The days of cheap, good data, however, may be coming to an end. Recently, some have discovered that a non-trivial proportion of the data collected on MTurk is "suspicious," generated either by "non-respondents" (bots) or non-serious respondents (e.g., Bai 2018; Dreyfuss 2018; Ryan 2018). This poses problems for those who rely on MTurk for survey and experimental research. If bots or survey satisficers provide more or less random answers, they could introduce noise that would bias average treatment effects toward zero.

We suspect, however, that threats to data quality on MTurk are potentially more grave. As we detail below, the nature of the platform offers `Workers`—experimental participants, for social science purposes—unique incentives to misrepresent themselves and their attitudes, beliefs, and preferences. Moreover, existing signals of quality are likely upwardly biased, making it difficult for `Requesters`—in our case, researchers—to distinguish between more conscientious `Workers` and those attempting to game the system. This ambiguity also means that MTurk may be particularly attractive to internet trolls who can reap (minor) financial gains while engaging in the same kind of humorous or provocative behavior they exhibit elsewhere online. To the extent that insincere responding is correlated with other variables of interest—for example, belief in political misinformation (e.g. Lopez and Hillygus 2018)—experimental treatment effects on such variables will be biased.

Spurred by these concerns, we fielded an original study in August 2018 to assess low-quality responding on MTurk and its 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 filed. We further collected IP-level metadata, such as the estimated location of the device, to more closely examine responses. Finally, we used survey completion times to identify potential survey satisficers and included a battery designed to indirectly assess "trolling" to determine how many `Workers` responded non-seriously.

We find that 11% of respondents circumvented location requirements or used multiple devices from the same IP address, while 16% of responses came from blacklisted IP addresses. Approximately 5–7% of respondents also engaged in trolling or satisficing. In all, about 25% of responses collected on MTurk appear untrustworthy.

Comparing the new survey to two studies conducted on MTurk in 2015, we find that the rate of low quality responding has increased anywhere from 50% to 300% over the past three years. And comparing the 2015 MTurk data to contemporary data collected via Survey Sampling International (since rebranded as Dynata), we show that MTurk is more likely to yield low-quality responding than other, more centrally managed online survey platforms.

Perhaps most importantly, we show that low-quality responses bias experimental results. Respondents who misrepresent themselves or troll differ from other survey-takers in how they respond to a vignette experiment embedded in our original survey. Specifically, these suspicious respondents attenuate treatment effects by introducing noise into the data; low-quality responses bias treatment effects downward by an average of roughly one percentage point, or 10% of our average treatment effect among non-suspicious respondents. This suggests that data collected on MTurk may be prone to yielding Type II errors.

While we find relatively low response quality, we believe a few changes in how we collect data on MTurk can improve things substantially. To that end, we conclude with a few recommendations.

2 Incentives For Quality on MTurk

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

Incentives for quality, however, vary by how hard it is to observe quality (Akerlof 1970). `Requesters`, for instance, often cannot directly observe `Workers'` demographic information or the location from which they are taking the survey, and `Workers` plausibly 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. `Workers` may also create multiple accounts and complete the same `HIT` multiple times, even when they are explicitly prohibited from completing each `HIT` more than once.

But these are just two examples—the problem is more general. MTurk was originally designed to be used internally at Amazon; humans performed simple classification tasks, like identifying patterns in images, that proved difficult for computers to complete (Pontin 2007). Mechanical tasks like these and others have a correct answer, and `Requesters` can track `Worker` quality by checking performance on known-knowns periodically or by comparing how often `Workers` agree with the majority of their peers (e.g., Garz et al. 2018).

With surveys, however, quality is nearly impossible to observe. Most social scientists use MTurk to solicit `Workers'` opinions, beliefs, and attitudes, which often lack an objectively correct answer. This makes it difficult to parse genuine responses from insincere ones. Except for cases in which a respondent takes extraordinarily little time to finish, researchers cannot accurately gauge whether or not participants are even reading the questions. Even selecting the first response option to multiple questions in a row is not conclusive evidence of satisficing (Krosnick, Narayan and Smith 1996; Vannette and Krosnick 2014). `Workers` could exploit this opacity by rushing through surveys to receive their compensation as quickly as possible.

While the concern applies to all survey platforms, the problem is likely worse on MTurk. MTurk, unlike other online survey platforms, lacks a standing relationship between respondents and those who curate samples. If the typical researcher uses MTurk two to three times a year, she has little incentive to sink resources into monitoring quality; instead, her investment is typically capped at the payout rate. On the other hand, survey vendors' business model is based upon providing high quality data to clients. Consequently, these firms have clearer incentives to monitor data quality, retaining well-behaved respondents and dropping those who raise red flags.

Moreover, the longstanding relationships that survey research firms build with their respondents afford them a clearer signal of respondent quality. As respondents take more surveys, they generate more data, which provides firms more opportunities to aggregate what might otherwise be individual weak signals into a more complete profile of respondent behavior. From there, vendors can choose to exclude poor-performing respondents from their subject pool. This in turn provides respondents with incentives to behave honestly, as they know their performance is being monitored by a large company that prizes data quality.

When it comes to MTurk, however, the only signal of `Worker` quality that `Requesters` can send to the market is `HIT` approval—that is, whether or not the `Worker` completed the task as assigned. MTurk tracks the percentage of `Workers`' completed `HITs`—of all kinds, not just surveys—as a signal of quality. While `HIT` completion rates may prove a useful signal for researchers using the platform to assess `Worker` performance on *objective* tasks, the difficulty of judging the quality of survey responses may preclude this metric's usefulness for social science research.

Worse, the `HIT` completion rate itself is likely upwardly biased, weakening any potential signal it sends. Not only is spot-checking data for response quality time consuming for `Requesters`, but treating sincere responses as insincere can be costly. `Workers` who are denied a payout can retaliate against `Requesters` by posting negative reviews on sites like Turkopticon, which provides `Workers` with detailed information about `Requesters`' average ratings and reviews of their `HITs`. Given these challenges—and the fact that the marginal cost of approving questionable work is typically only a few cents—`Requesters` often batch approve completed `HITs`, making the `HIT` completion metric a biased signal of `Worker` quality.

Given this information asymmetry, `Workers` have strong incentives to game the system by misrepresenting where are they are located, masquerade as someone else to "double dip," and complete surveys insincerely or inattentively.[1] The difficulty in assessing response quality also means MTurk may be particularly attractive to people who enjoy trolling—i.e., providing outrageous or misleading responses—as it allows them to make money while indulging their id (e.g., Cornell et al. 2012; Lopez and Hillygus 2018; Robinson-Cimpian 2014; Savin-Williams and Joyner 2014).

All of this suggests that data collected on MTurk may not be of as high quality as researchers often assume. There are distinct incentives for `Workers` to misrepresent themselves, and existing signals of `Worker` quality may not capture the degree to which `Workers` engage in bad behavior. Consequently, low quality responses on MTurk may be far more common than is typically assumed.

[1]Some `Workers` may even use software to autofill forms. Examples of these kinds of programs can be found here or here.

3 Assessing the Quality of Responses on MTurk

To investigate data quality, we posted a survey on MTurk 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. `Workers` were told the survey would take about 10 minutes to complete, and we paid $0.60 for each completed `HIT`. In keeping with best practices (Peer, Vosgerau and Acquisti 2014)—and, thus, consistent practices, for external validity—we restricted participation to MTurk `Workers` with a `HIT` completion rate of at least 95%.

First, to assess how many `Workers` are using form-filling software or bots to complete surveys quickly, we used `No CAPTCHA reCAPTCHA` (Shet 2014), which uses mouse movements to estimate whether activity on the screen is produced by a human or a computer program.

Bots are only one potential source of low quality data on MTurk. To identify people who masquerade as someone else or provide misleading answers about their location, we exploited data on IP addresses.[2] First, 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 pull data on IP addresses from multiple services. In particular, Know Your IP uses MaxMind (MaxMind 2006), the largest, most trusted provider of geoIP data, to provide locations of the IP addresses. Know Your IP also collects data on blacklisted IP addresses,[3] which often appear on the same traffic anonymization services that people use to evade location filters. Know Your IP pulls blacklist data from ipvoid.com, which collates

[2]While IP addresses are not permanent, the turnover rate is low. Accordingly, temporally proximal inferences on IPs are reasonably reliable.

[3]IP addresses are blacklisted for two main reasons: (1) a website associated with the IP is caught spreading malware or engaging in phishing, (2) bad Internet traffic like a DDoS attack originates from the IP.

data from 96 separate blacklists.

We also collected information about how many responses originated from the same IP address. This information is useful because only devices that share the same router—or Virtual Private Network/Virtual Private Server—can have the same IP address. At minimum, this tells us how many responses originate from the same organization or household or which IPs used traffic anonymization software. Multiple `HITs` completed from the same IP address could reflect participation from several individuals (such as members of a family), but given current incentive structures, we suspect at least some of these data points reflect cases where individuals used multiple accounts to complete the same `HIT` more than once.

While we cannot identify all survey satisficers, one might reasonably assert that `Workers` who completed the survey extraordinarily quickly may not have provided meaningful responses. To that end, we recorded and examined response times. Our median completion time was 573 seconds—or nine minutes and 33 seconds, 27 seconds under the ten minute target we provided. We flagged respondents as outliers if they finished 167% outside the interquartile range (IQR) of completion times. Accordingly, fast outliers were those who completed the survey in 245 seconds (four minutes and 5 seconds) or less. Slow outliers were those who completed the survey in 1,139 seconds (18 minutes and 59 seconds) or more.

To identify "trolls" and other non-serious respondents, we followed Lopez and Hillygus (2018) in asking a series of "low incidence screener" questions about rare afflictions, behaviors, and traits (Cornell et al. 2012; Robinson-Cimpian 2014; Savin-Williams and Joyner 2014). Specifically, we asked respondents whether they or an immediate family member belonged to a gang, whether they had an artificial limb, whether they were blind or had impaired vision, and whether they had a hearing impairment. We also asked respondents how much they slept. We coded anyone reporting sleeping more than ten hours or fewer than four hours as unusual. In keeping with previous research, we flag respondents as satisficing or trolling if they presented low-incidence characteristics on two or more of these items (Lopez and Hillygus 2018).[4] At the end of the survey, we also asked respondents an explicit question about how sincerely they respond to surveys. We compare responses to this question with responses to the screener questions to assess respondent honesty. (For detailed question wording, see SI 1.1.)

Results

We start by looking at evidence for the use of bots. 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 could not record the IP address as suspect.) Of the 1,991 responses, 106 responses came from an IP that appears in our dataset more than once (see Table SI 1.1). As noted previously, this could be because multiple people in the same household completed the `HIT`, but the more plausible explanation is that respondents used multiple accounts to submit the same `HIT` multiple times.[5]

A majority of responses (1,866) originated from within the United States (see Table 1). Of the 125 foreign responses, 42 were from Venezuela and 17 were from India. (See Table SI 1.2 for a complete 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.

[4]It is plausible, even likely, that people with physical disabilities or those that come from marginalized groups are overrepresented on MTurk. Ideally, we would have more defensible priors than the naïve comparisons we present below.

[5]Even if multiple people from the same household completed the survey, knowing this would still be useful for survey researchers, as it would affect standard error calculations.

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. Similarly, the distribution of cities from which responses were filed suggests irregularities consistent with contemporaneous assessments of MTurk data quality (Kennedy et al. 2018; Ryan 2018) (see Table SI 1.3).

Table 1: *Frequency of Different Types of Suspicious IPs*

Type of Suspicious IP	n
Missing	9
Blacklisted	321
Duplicated	106
Foreign	121
Any of the Above	406

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.

We also examined how many `Workers` may have engaged in satisficing when completing our survey. We found that just under 2% of respondents were "fast outliers" who completed the survey in under 245 seconds. Consistent with folk wisdom, far more respondents (14.8%) were classified as "slow outliers."[6]

Next, we examined the frequency of insincere or 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 and 14.5 times their respective rates in the population.[7] These large deviations from the national norm are

[6]A longstanding rule for designing MTurk `HITs` has been to give Turkers far longer to complete the task than necessary, as their attention may be drawn away from the computer.

[7]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).

possible but unlikely. Questions on gang membership have similarly implausible numbers, with about 6% of respondents reporting having a family member in a gang—compared to a rate of about half a percent in the overall population (National Gang Intelligence Center (U.S.) 2012). 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,we classify 125 respondents (roughly 6%) as "trolls" accordingly.

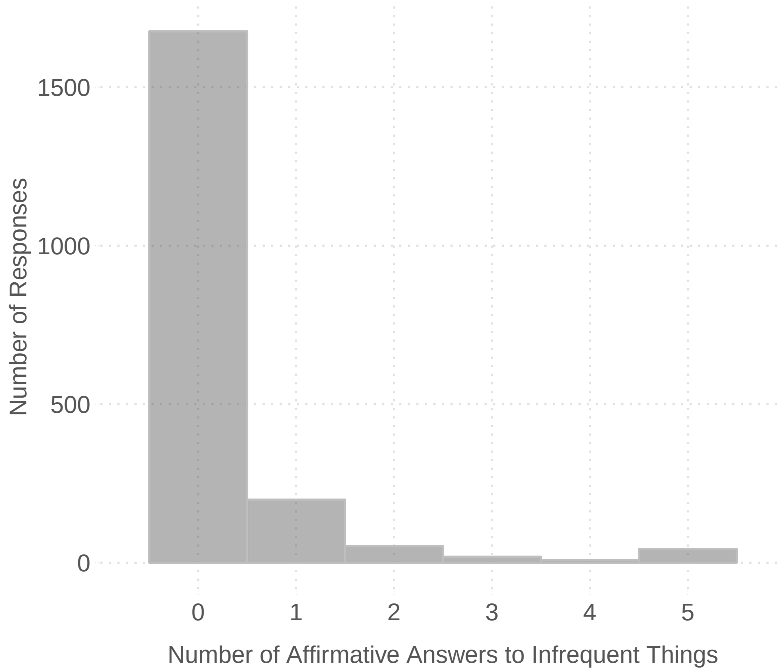
Table 2: *Respondents Reporting Rare Behaviors/Traits*

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

Additionally, 99 respondents (or roughly 5% of the sample) reported that they "always" or "almost always" provided humorous or insincere responses to survey questions. These respondents were more likely to be classified as trolls, suggesting that the low-incidence screeners identify insincere responding and not just inattentiveness. Of the 1,875 respondents who responded affirmatively to one or fewer low-incidence screeners, nearly 93% reported that they "never" or "rarely" answered humorously or insincerely. By contrast, roughly 58% of the 125 classified as trolls said that they usually answered sincerely ($\chi^2 = 179.0, p < 0.001$). In all, 5–7% of `Workers` recruited for this study showed insincere response tendencies.

To assess relationships between various indicators of low-quality responding, IP address data with our measure of trolling based on the low-incidence screeners. 38 of the 408 responses from "bad" IPs (about 9% of the sample) replied in the affirmative on two or more of these items, compared to just under 6% of the remaining responses. This difference is sta-

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



tistically significant but not eye-catchingly large. But neither did we expect it to be: people who game the MTurk system want to do enough to get paid while flying under Amazon's radar. Whether we want data from these actors, howeer, is another question.

Surprisingly, we find that potential trolls and potentially fraudulent IP addresses take significantly longer on the survey on average (by 146 seconds, $p < .001$) and are significantly more likely to be slow outliers ($\hat{\beta} = 0.13$, $p < .001$). On the other hand, they are no less likely to be fast outliers ($\hat{\beta} = -0.00$, $p = .79$). We therefore do not count fast outliers as untrustworthy responses. (And, as we show in the next section, unlike other flagged respondents, these speedsters do not appear to provide lower-quality data.)

In all, 495 responses are from IPs that are duplicated, located in a foreign country, or blacklisted, or provided affirmative answers to two or more of the low-incidence questions. Altogether, nearly a quarter of responses are potentially untrustworthy.[8]

[8]Though this figure seems rather high, we suspect that it may in fact underestimate

4 Low Quality Responses Across Time & Platforms

We have argued that low quality responding may be more prevalent on MTurk, due to its unique incentive structure, than other, more centrally managed survey platforms. Further, we have posited that MTurk's incentive structure will result in increasingly low data quality as the platform attracts more bad faith `Workers` who attempt to game the system for minimal effort and maximum payout.

Thus far, we have only theorized about these trends; here, we provide evidence. To do so, we exploit IP data collected from three other studies—two conducted on MTurk and one conducted using Survey Sampling International (SSI). The first MTurk study, fielded as part of Ahler and Goggin (2019) (`AG`), collected responses from February 3-19, 2015; the second, carried out from March 9-May 6, 2015, appears as part of Ahler and Broockman (2018) (`AB`). Our final dataset consists of the 2015 IGS-California Poll omnibus survey, fielded on Survey Sampling International (now known as Dynata)'s platform from August 11-26. As all three studies were conducted in 2015, we are limited in our ability to draw conclusions about yearly trends in low quality responding on MTurk. That said, investigating these samples' data quality can shed light on whether suspicious behavior on MTurk has increased over the past three years.

As before, we made use of Know Your IP (Laohaprapanon and Sood 2018) to de-

---

the prevalence of some types of low-quality responding. Individuals pay shockingly little attention to online surveys while completing them (e.g., Woon 2017); Mummolo and Peterson (2019) found that only about 35-50% of participants passed a manipulation check (Appendix B). With few incentives for survey respondents to carefully read and process every question, we believe it is quite likely that many of the `Workers` recruited for our study also failed to pay attention to portions of our survey; given our current data, however, there is no way to know for certain.

termine the number of respondents in each study who circumvented location requirements, completed the survey more than once, took the survey from a location outside the U.S., or completed the survey using a blacklisted IP address. Table 3 shows the frequency of different types of low quality responding by study.[9] For comparison, we include similar statistics from our August 2018 survey.

Unfortunately, low-quality responding on MTurk appears not to be a recent phenomenon. Perhaps more troubling, rates of suspicious responding on MTurk appear to have grown substantially over the past three years. Our 2018 study has nearly four times the proportion of low quality responses as `AG` (6%) and nearly double the proportion of low quality responding as `AB` (13%). In addition, the2015 data (analyzed in 2019) likely underestimates the proportion of IP addresses that are problematic. IP addresses turn over, especially those flagged for suspicious behavior. The data from 2015, therefore, may have a slight positive bias: some blacklisted IP addresses have likely been reassigned since then, which underestimates the scope of the problem at the time of data collection. In any case, while our results here are by no means conclusive, they are consistent with declining MTurk data quality.

Table 3: Prevalence of Low-Quality Responding by Study

Survey	Platform	Missing	Blacklisted	Duplicated	Foreign	Any	N	% Low Quality
August 2018 Study	MTurk	9	321	106	121	406	2,000	20.30%
Ahler & Broockman (2015)	MTurk	0	23	182	58	257	2,045	12.57%
Ahler & Goggin (2015)	MTurk	0	5	28	20	51	898	5.68%
UC Berkeley IGS Poll (2015)	SSI	0	18	72	7	95	2,295	4.14%

Our results also suggest that low-quality data may be a larger problem on MTurk than other online survey platforms. Roughly 4% of the data in the SSI-administered IGS

[9] As these 2015 studies did not include any questions designed to detect trolling or satisficing, we only include statistics related to suspicious IPs here.

Poll is of low quality, only slightly smaller than the proportion in `AG`. That being said, this figure is less than a quarter of the proportion of low-quality responses in `AB` and about a fifth of the proportion of low quality data in our August 2018 study. While it is conceivable that low quality responses on SSI/Dynata and other online platforms have also increased in the past few years, our results here—and the incentive structures within MTurk—suggest that data collected on MTurk is likely to be of poorer quality than data gathered through platforms owned by firms with incentives to police quality.

5 Consequences of Low Quality Responding

The results above suggest that there are at least three significant concerns with survey data collected on MTurk. First, a sizeable proportion of respondents took the survey from outside the United States. If—as we suspect—the majority of these respondents are foreigners, many of our responses are provided from people from outside the sampling frame. Second, a significant number of respondents filed multiple responses. Finally, a non-trivial proportion of people appeared to intentionally respond humorously to the survey.

Some might assert that these are annoyances—but not fatal to research—because they "merely" add noise to data. For example, if `Workers` respond "randomly" by rushing through the survey—or if foreign `Workers` provide random answers because they do not understand English or American politics—they introduce noise. But this noise itself may be a bigger problem than many assume: it attenuates correlations and can bias estimates of frequencies and means on some variables. For instance, even answering questions randomly can positively bias estimates of how many people know something (Cor and Sood 2016).

Trolling presents potentially graver consequences. If people respond humorously or with the aim of being provocative, they will instead introduce more *systematic* error into estimates of the prevalence of certain attitudes (e.g., Lopez and Hillygus 2018). In either case—attenuation bias or systematic error from trolling—these errors threaten our ability to draw accurate inferences.

To study how low-quality responses influence the substantive conclusions reached in a study, we embedded an experiment on partisan stereotyping into the August 2018 survey. We replicated a study from Ahler and Sood (2017), examining the degree to which people rely on the representativeness heuristic when making judgments about party composition. Specifically, the study investigates the degree to which people use information about how social groups "sort into" one of the two parties (at the expense of other relevant considerations) to make inferences about aggregate party composition. One way to assess this—specifically, the "at the expense of other relevant considerations" part—is to exploit the *conjunction fallacy*, a cognitive error that occurs when people assert the probability of two events occurring together is greater than the probability of either occurring separately (Tversky and Kahneman 1974).

Ahler and Sood (2017) itself is a modification of Tversky and Kahneman's (1974) "Linda Problem," which presented respondents with the following question:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable?
>
> - Linda is a bank teller.
> - Linda is a bank teller and is active in the feminist movement.

The latter option is logically impossible, as the probability that Linda is both a bank teller *and* active in the feminist movement will always be less than or equal to the probability that Linda is a bank teller. Therefore, when respondents select the second option, they commit the conjunction fallacy as a result of their overreliance on representative characteristics.

Ahler and Sood (2017) modified the Linda problem by manipulating the characteristics of the target in the vignette (i.e., making the character more or less representative of one of the two parties) to assess which characteristics people weigh most heavily in party stereotypes (Ahler and Sood 2018). To do so, they introduced respondents to a character named James, randomly and independently manipulating particular party-representative characteristics (like gender, race, sexual orientation, and religion) 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. 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 the vignette, we 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." In selecting option (2) or (3), respondents commit the conjunction fallacy. In their original study, Ahler and Sood (2017) found, unsurprisingly, that exposure to characteristics that are representative of the Democratic (Republican) Party leads individuals to commit the Democratic (Republican) conjunction fallacy. By including a replication in the present survey, we can examine whether suspicious respondents react differently than traditional survey-takers to an already-validated treatment.

To determine if and how low-quality responses moderate treatment effects, we estimated the *average marginal component effect* (AMCE) of each independently randomized characteristic interacted with an indicator for a low-quality response 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 (omitting one value per variable) to analyze the data. Thus, our model takes the following form, with i indexing respondents and j indexing possible values of the dependent variable:

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

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:

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

where X_k denotes our vector of randomly and independently assigned characteristics of James (his race, sexuality, etc.) and LQ_i is an indicator for `low quality` response. We operationalize `low quality responses` three ways in three different models: first as all respondents flagged for any reason, then as duplicated/blacklisted IP addresses, and finally as respondents flagged for potential trolling.

Full model results are available in SI 1.2. For ease of interpretation, we present marginal effects in Table 4, specified as the change in the predicted probability of committing the Democratic/Republican conjunction fallacy. We first present results for all non-flagged respondents (column 1) and then all low-quality respondents (duplicated/blacklisted IP addresses and respondents we suspect are non-serious (column 2). Finally, we present the results for flagged IP addresses alone (column 3) and potential trolls 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 described as black, gay, secular, or as having liberal policy preferences; they are also more likely to commit the Republican conjunction fallacy when James is presented as evangelical or 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 react differently. AMCEs are generally attenuated among respondents flagged for any reason. The magnitude of this difference 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. Oddly, the effect of the conservative cue is substantively larger among suspicious respondents, but this difference from non-suspicious respondents is not precisely estimated.

Averaging these differences in treatment effects (weighted inversely by their estimated standard errors) yields a difference in average treatment effects between suspicious and non-suspicious respondents of 3.7 percentage points (95% confidence interval (CI): [0.10, 6.5]). When we calculate a precision-weighted average difference between treatment effects in the entire sample and those among non-suspicious respondents, we observe an attenuation effect of roughly 0.9 percentage points [95% CI: [0.3, 1.6]). We can contextualize this attenuation

Table 4: Impact of Low-Quality Responding on Treatment Effects - Marginal Effects

	Non-flagged respondents ($n = 1,507$)		All low-quality respondents ($n = 484$)		Flagged IPs only ($n = 359$)		Non-serious respondents only ($n = 87$)	
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)	**13.9%**	**-9.7%**	*5.1%*	*-3.7%*	**8.1%**	**-5.9%**	*-5.9%*	*5.0%*
Gay (vs. straight)	**19.1%**	**-13.2%**	***9.3%***	***-6.8%***	**12.1%**	**-8.8%**	*-2.9%*	*2.5%*
Evangelical (vs. nothing)	**-5.8%**	**4.2%**	0.0%	0.0%	-1.4%	1.0%	12.4%	-10.3%
Secular (vs. nothing)	**6.6%**	**-4.5%**	7.3%	-5.2%	5.5%	-4.0%	17.8%	-14.1%
Liberal (vs. nothing)	**9.4%**	**-6.4%**	-0.9%	0.7%	2.6%	-1.9%	***-12.6%***	***11.7%***
Conservative (vs. nothing)	**-8.3%**	**5.9%**	**-11.1%**	**8.5%**	**-16.7%**	**12.7%**	-13.6%	12.2%

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 ($p < 0.1$).

effect by putting it in percentage terms: the observed precision-weighted average treatment effect among non-suspicious respondents is 8.9 percentage points, and the presence of suspicious respondents (and their noisy data) attenuates this estimated effect by 10.1% (see SI 1.3 for more on this estimation procedure).

Estimates are generally attenuated among responses with flagged IPs (column 3), but we find more puzzling results among trolls or satisficers (column 4). These potentially non-serious respondents were significantly more likely to profess James to be a *Democratic* salesman when James was described as evangelical, and more likely to commit the *Republican* conjunction fallacy when James had liberal views. Oddly, however, the effects of the secular and conservative cues were substantively large within this group—larger than those observed for non-suspicious respondents—and in the correct direction, albeit imprecisely estimated because of the small number of potential trolls. While potential trolls 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—and these results do not allow us to rule that possibility out.

Finally, we consider whether extraordinarily fast completion times produce lower-quality data and attenuate treatment effects. Contrary to conventional wisdom, they do not appear to do so. Fast outliers were 3.4 percentage points less likely to commit the conjunction fallacy, but this apparent difference is imprecisely estimated (95% CI: [-0.17, 0.10]) and relatively small (74% of respondents did so). Furthermore, as SI 1.4 shows, these fast outliers respond to the experimental treatments similarly to non-suspicious respondents in terms of their predicted probabilities of committing the party-particular conjunction fallacies.

One reason for this could be that people who complete surveys more quickly are better readers and comprehend survey material more quickly. We find that respondents who completed college—our best proxy for reading comprehension—are indeed 1.8 percentage points more likely be fast outliers ($p = 0.01$) and do complete the survey more quickly (albeit by just 19 seconds, $p = 0.06$). We find it more likely, however, that people classified as fast outliers simply take lots of surveys and are better at automatically processing the information they contain. This, of course, yields its own data quality problems (e.g., Huff and Tingley 2015). But if fast response is a function of taking many surveys, at least in this case, high-volume respondents reacted to treatments similarly to other respondents.

6 Discussion and Conclusion

Our study demonstrates that there are significant data quality problems on MTurk. About a quarter of our data is potentially untrustworthy, and "problematic" respondents on the platform respond differently to experimental treatments than other subjects. Specifically, we find that bad behavior (in the form of cheating or trolling) adds noise to the data, which attenuates treatment effects—in our case, by 10%.

Current data quality may be poor, but what's the prognosis? Since concerns about a "bot panic" surfaced in the summer of 2018, Amazon has implemented several reforms designed to cut down on the number of `Workers` gaming the platform for personal gain. These measures include requiring U.S. `Workers` to provide official forms of identification, shutting down sites where `Worker` accounts are traded, and monitoring `Workers` using IP network analysis and device fingerprinting (Amazon Mechanical Turk 2019). While these measures may catch some of the worst offenders, we believe that given the platform's strategic incentives, `Worker` quality will continue to decline as bad actors devise new ways to game the system.

Unless we can craft and implement better methods to assess and incentivize quality responding, the chances of improvement 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 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 from the same IP.
- Include questions to detecting trolling and satisficing but do not copy and paste from a standard canon, which 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 responses in this analysis. But for the bulk of MTurk studies (i.e., those not being done to audit the platform), 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 `Workers`, and (2) it should not be easily gamed.

  Rather than withhold payments, a better policy may be to implement quaity filters and let `Workers` know in advance that they will receive a bonus payment if their work is completed honestly and thoughtfully. This would lead to a weak signal propagating the market in which 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 `Requesters'` generosity, fairness, etc.
- Be mindful of compensation rates. While stingy wages will lead to slow data collection times and potentially less effort by `Workers`, unusually high wages may give rise to adverse selection—especially because `HITs` are shared on Turkopticon, etc. soon after

posting. A survey with an unusually high wage gives large incentives to foreign `Workers` to try to game the system despite being outside the sample frame. Social scientists who conduct research on MTurk should stay apprised of the current “fair wage” on MTurk and adhere accordingly.

- Use `Worker` qualifications on MTurk and include only `Workers` who have a high percentage of approved `HITs` into your sample. While we have posited that `HIT` completion rates are likely a biased signal for quality, filtering `Workers` on an upper-90s completion rate may weed out the worst offenders. Over time, this may also change the market.

This problem may not be limited to MTurk. Indeed, we found evidence of suspicious responding in a 2015 sample curated by SSI. However, MTurk is likely 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. On any paid platform, non-serious responding is bound to be a concern, but our analysis suggests the problem is magnified on MTurk—and, moreover, has likely worsened as MTurk has become more widely used in the social sciences.

We conclude by noting that issues with data quality on platforms like MTurk may necessitate a reconsideration of the relationship between social scientists and our human subjects. 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 our call for respect for data is inconsistent with respect for persons, beneficence, and justice. By following the aforementioned guidelines—and being clear about the 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 David Broockman. 2018. "The Delegate Paradox: Why Polarized Politicians Can Represent Citizens Best." *Journal of Politics* 80(4):1117–1133.

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

Ahler, Douglas J. and Gaurav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *ournal of Politics* 80(3):964–981.

Ahler, Douglas J. and Stephen N. Goggin. 2019. How Does One Recognize #FakeNews? Assessing Competing Explanations Using a Conjoint Experiment. In *Annual Meeting of the Midwest Political Science Association*. Chicago: .

Akerlof, George A. 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84(3):488–500.

Amazon Mechanical Turk. 2019. "MTurk Worker Quality and Identity." Available at https://blog.mturk.com/mturk-worker-identity-and-task-quality-d3be46d83d0d.

Bai, Hui. 2018. "Evidence that a Large Amount of Low Quality Responses on MTurk Can be Detected with Repeated GPS Coordinates." Available at 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. Available at 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.

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.

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).

Institute of Governmental Studies at the University of California, Berkeley. 2015. "2015 Omnibus Survey." https://www.igs.berkeley.edu/igs-poll/berkeley-igs-poll.

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. Available at 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.”. Available at 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*. Chicago: .

MaxMind, LLC. 2006. “GeoIP.”. Available at https://www.maxmind.com/en/home.

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.

Mummolo, Jonathan and Erik Peterson. 2019. “Demand Effects in Survey Experiments: An Empirical Assessment.” *American Political Science Review* 113(2):517–529.

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.

Pontin, Jason. 2007. "Artificial Intelligence, With Help From the Humans." *The New York Times* 25 March. Available at https://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html.

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." Available at 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'". Available at https://security.googleblog.com/2014/12/are-you-robot-introducing-no-captcha.html.

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.

Woon, Jonathan. 2017. "Political Lie Detection.". Unpublished manuscript, available at https://rubenson.org/wp-content/uploads/2017/11/woon.pdf.

SI 1 Supporting Information

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

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

Table SI 1.2: *Response Country of Origin*

Country	Freq
United States	1,870
Venezuela	42
India	17
Canada	6
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
Czech Republic	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	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

SI 1.1 Question Wording Text

Experimental Manipulation

Please read the descriptions of recent college graduates on this screen and the next and answer the related questions.

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`).

GPA Guess

What do you think James' GPA was in college?

- 3.80 - 4.00
- 3.50 - 3.79
- 3.00 - 3.49
- 2.50 - 2.99
- 2.49 or below

Conjunction Fallacy

Which of the following do you think is most likely?

- James works in sales
- James works in sales and is an active supporter of the Democratic Party
- James works in sales and is an active supporter of the Republican Party

Low Incidence Screener 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

Honesty Self-Report

Finally, we sometimes find people don't always take surveys seriously, instead of 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 Low-Quality Responses on Treatment Effects - Full Ordered Logit

	All respondents	Suspicious IPs	Non-serious respondents
Low-quality response	-0.15	-0.31	-0.28
	(0.26)	(0.29)	(0.59)
Black	-0.62	-0.61	-0.61
	(0.10)	(0.10)	(0.10)
Black * LQ	0.41	0.28	0.85
	(0.20)	(0.23)	(0.42)
Gay	-0.83	-0.83	-0.82
	(0.10)	(0.10)	(0.10)
Gay * LQ	0.46	0.34	0.95
	(0.20)	(0.22)	(0.42)
Evangelical	0.26	0.26	0.26
	(0.12)	(0.12)	(0.12)
Evang. * LQ	-0.31	-0.24	-0.77
	(0.24)	(0.27)	(0.54)
Atheist/agnostic	-0.31	-0.29	-0.29
	(0.24)	(0.13)	(0.13)
AA * LQ	0.00	0.06	-0.42
	(0.25)	(0.28)	(0.53)
Liberal	-0.42	-0.41	-0.41
	(0.13)	(0.13)	(0.13)
Lib. * LQ	0.31	0.31	0.95
	(0.24)	(0.27)	(0.51)
Conservative	0.36	0.36	0.36
	(0.12)	(0.12)	(0.12)
Con. * LQ	0.10	0.09	0.21
	(0.24)	(0.27)	(0.51)
Cut 1	-0.60	-0.59	-0.59
	(0.13)	(0.13)	(0.13)
Cut 2	0.67	0.65	0.65
	(0.13)	(0.13)	(0.13)
Pseudo R^2	0.04	0.05	0.05
n	1,991	1,866	1,594

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

SI 1.3 Calculating Attenuation Effects

From the data and the ordered logistic regression model specified in the text, we estimate the average change in respondents' predicted probability of committing the Democratic and Republican conjunction fallacies when they see that James has k_1 attribute instead of some omitted category k_0. (For example, k could be race, with k_1 meaning that James is black and k_0 that he is white.)

We estimate these average changes in the effect of attributes k among: (1) the full sample, (2) non-suspicious respondents, and (3) suspicious respondents. From there, we calculate the average difference in treatment effects, weighted inversely by the standard errors of those estimated differences, between pairs of these three groups. The difference between groups 1 and 2 is the average attenuation effect in percentage point terms. We can further contextualize this difference by dividing the estimated effects of k in group 1 by the estimated effects in group 2, which yields the relative size of the observed effect to the "real" effect (i.e., the effect among non-suspicious respondents only)—the *attenuation ratio*. We calculate an average attenuation ratio, weighted again by the inverse of the standard error of these estimated differences. Subtracting the attenuation ratio from 1 yields the attenuation effect in percentage point terms.

SI 1.4 Do Speedy Respondents Produce Low-Quality Data?

Contrary to conventional wisdom, we do not find that respondents who are extraordinarily fast in their completion of the survey provide low-quality data. At the very least, modeling response to the James problem as a function of the experimental treatments, being a fast outlier, and the interaction of the treatments with fast-outlier status, we find that speedy respondents react to our experimental treatments quite similarly to respondents who are neither extraordinarily speedy or slow. In only one out of six cases do they appear to respond significantly differently—the atheist/agnostic cue ($p = .09$)—but the coefficient is incorrectly signed for our hypothesis; fast outliers are slightly more responsive to this cue than slower non-suspicious respondents are. (Note that this analysis is limited to respondents who are not otherwise "suspicious" aside from responding quickly.)

Table SI 1.5: Impact of Fast Completion Times on Treatment Effects - Full Ordered Logit

	DV: James Experiment
Fast outlier	0.55
	(1.04)
Black	-0.62***
	(0.10)
Black * fast	0.97
	(0.97)
Gay	-0.83***
	(0.10)
Gay * fast	-0.13
	(0.86)
Evangelical	0.26**
	(0.12)
Evang. * fast	-0.52
	(1.00)
Atheist/agnostic	-0.26
	(0.13)
AA * fast	-1.84*
	(1.08)
Liberal	-0.41***
	(0.13)
Lib. * fast	-0.55
	(1.06)
Conservative	0.37
	(0.12)
Con. * fast	-0.99
	(0.96)
Cut 1	-0.57
	(0.14)
Cut 2	0.65
	(0.14)
Pseudo R^2	0.05
n	1,507