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

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

January 22, 2019

Abstract

Amazon's Mechanical Turk 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) or 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 engaged in suspicious behavior. About 20% of respondents either circumvented location requirements or took the survey multiple times. In addition, 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. Expectedly, we find response quality impacts experimental treatments. On average, low quality responses attenuate treatment effects by approximately 9%. 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, and John Sides for the useful comments and suggestions.

[†]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

Amazon's Mechanical Turk (MTurk) has rejuvenated the social sciences. It has freed researchers from reliance on the "narrow database" of social science undergraduates (Sears 1986) and dramatically reduced the cost and inconvenience of fielding a study (e.g., Berinsky, Huber and Lenz 2012; Casler, Bickel and Hackett 2013; Paolacco and Chandler 2014). Over the past few years, the platform has become a popular venue for social science experimentation, and for good reason. Though MTurk samples may not be representative of a broader population, random assignment to experimental treatments eliminates selection bias (Gerber and Green 2012; Shadish, Cook and Campbell 2002), making the platform a convenient means for testing causal hypotheses. Survey respondents on MTurk 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 study 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 nationally-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 researchers 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 to those who rely on MTurk for experimental research. If bots or survey satisficers provide more or less random answers to questions, they could introduce noise that biases average treatment effects toward zero.

We suspect, however, that threats to data quality on MTurk are much more widespread than commonly understood. As we detail below, the nature of the platform offers MTurk `Workers`—experimental participants, for social science purposes—unique incentives to mis-

represent themselves and their attitudes. 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 humorous or provocative responding correlates with other variables of interest—for example, belief in political misinformation (e.g. Lopez and Hillygus 2018)—experimental treatment effects may also be biased in other ways.

Spurred by these concerns, we fielded an original study in August 2018 to gauge the prevalence of several types of low-quality responses on MTurk and their 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, to more closely examine suspicious responses. In addition, we examined survey completion times to identify potential survey satisficers. Finally, we included a battery of questions designed to measure satisficing or trolling to determine how many `Workers` responded non-seriously 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 at least 5–7% of respondents engaged in trolling or satisficing. In all, we find that about 25% of the responses are potentially untrustworthy.

We also find that these low-quality responses bias experimental results. Using a vignette experiment embedded in the same survey, we find that respondents who misrepresent themselves or troll differ from other survey-takers in how they respond to treatments. Specifically, these suspicious respondents attenuate treatment effects by introducing significant

noise into the data; low-quality responses bias treatment effects downward by an average of roughly one percentage point, or 8.6% of our average treatment effect among non-suspicious respondents. This suggests that low-quality data may be responsible for Type II errors in studies conducted using data from MTurk.

While our data suggest that the quality of responses on MTurk is low, we believe a few changes to data collection procedures can increase quality substantially. To that end, we conclude with a few recommendations for researchers about how to increase response quality in future work.

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. `Requesters` 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 1970). `Requesters`, for instance, often cannot directly observe a `Worker's` demographic information or the location from which s/he is taking the survey. And `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. Or `Workers` may 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 for internal use at Amazon; human workers were recruited to perform simple clas-

sification 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 individual Workers agree with the majority of their peers (e.g., Garz et al. 2018).

When it comes to surveys, however, quality of work is nearly impossible to observe. Most social scientists use MTurk to solicit Workers' opinions, beliefs, and attitudes, which by definition have no right answer. This makes it difficult to parse genuine responses from insincere ones. Except for cases where a respondent takes extraordinarily little time to finish a survey, researchers cannot accurately gauge whether or not a respondent is even reading the 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). Workers could exploit this opacity by rushing through the survey to complete it—and thereby receive their payout—as quickly as possible.

While the concern applies to all survey platforms, the problem is likely worse on MTurk. First, unlike other survey platforms, on MTurk, there is typically no standing relationship between Workers and Requesters. Lack of a long-term relationship means that Requesters have few incentives to sink resources into monitoring quality.

Second, the only signal of Worker quality that Requesters can send to the market is HIT approval—that is, whether or not they choose to pay the Worker after the task is completed. MTurk tracks the percentage of Workers' completed HITs—of all kinds, not just surveys—as a means to assess performance. While HIT completion rates may prove a useful signal for researchers using the platform to assess Worker performance on *objective* tasks, because of the previously discussed difficulties in judging the quality of survey response, these ratings may not be as effective of a means of discerning Worker quality for social science research.

Moreover, the HIT completion rate signal itself is likely weak and upwardly biased. Not only is spot-checking data for quality responses a time consuming task for Requesters, there are also high costs associated with committing a false positive—that is, treating sincere responses as insincere. 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 noisy signal of Worker quality.

Given the information asymmetry between Requesters and Workers, Workers have strong incentives to game the system by misrepresenting where are they are located, masquerading as someone else when "double dipping," or completing surveys insincerely or inattentively.[1] The difficulty in assessing quality responses also means MTurk may be particularly attractive to people who enjoy trolling—that is, providing outrageous or misleading responses—as it allows them to make money while indulging their id (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 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 responses on MTurk may be far more prevalent than is commonly understood.

[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 study how common these types of problematic responses are on MTurk, 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`, making our hourly pay rate about $3.60. In keeping with best practices (Peer, Vosgerau and Acquisti 2014), we restricted participation to MTurk `Workers` with a `HIT` completion rate of at least 95%.[2]

First, in an effort to assess how many `Workers` may be using automated software or bots to complete surveys quickly, we used `No CAPTCHA reCAPTCHA` (Shet 2014). `No CAPTCHA reCAPTCHA` uses mouse movements on the screen to estimate whether activity on the screen originates from a human or from a computer program.

As noted previously, bots are only one potential source of low-quality responses on MTurk. To identify people who masquerade as someone else or provide misleading answers about their location, we exploited data on IP addresses.[3] 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 multiple services that provide data on IP addresses. In particular, Know Your IP uses estimated locations of IP addresses retrieved from MaxMind (MaxMind 2006), the largest, most trusted provider of geoIP data. Know

[2]As noted above, though `HIT` completion rates are likely a noisy signal of `Worker` quality, we employ the filter here as an extra precaution.

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

Your IP also collects data on blacklisted IP addresses,[4] blacklisted addresses often appear on the same traffic anonymization services that others use to evade location filters. Know Your IP pulls data about where an IP is blacklisted from ipvoid.com, which collates 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 information tells us how many responses originate from the same organization or household. 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 multiple responses reflect cases where individuals used multiple accounts to complete the same `HIT` more than once.

To identify possible survey satisficers, we examined survey completion times. While we cannot identify all individuals who may have engaged in satisficing while completing the survey, one might reasonably assert that those `Workers` who completed the survey extraordinarily quickly may not have provided meaningful responses. Our average completion time was 573 seconds—or nine minutes and 33 seconds, 27 seconds under the ten minute target we gave respondents. We flagged respondents as outliers based on time 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 1139 seconds (18 minutes and 59 seconds) or more.

To identify "trolls" and other non-serious respondents, we followed Lopez and Hillygus

[4]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.

(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 themselves 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 for 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 answered "yes" or reported unusual behavior on two or more of these questions (Lopez and Hillygus 2018).[5] 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 get a measure of 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

[5]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. We provide data in the results section to describe how unlikely some of the traits, behaviors, and afflictions are in the general population.

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). As noted previously, this could be because multiple people in the same household completed the `HIT`, but we think the more plausible explanation is that respondents used multiple accounts to submit the same `HIT` multiple times.[6]

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 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. 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. Like Ryan (2018), we find that the city from which 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

[6]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.

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 completed the survey in under 245 seconds, or were classified as "fast outliers." Consistent with folk wisdom on MTurk, far more respondents (14.8%) were classified as "slow outliers"—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]

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 times and 14.5 times the respective rates in the population.[8] These large deviations from the national norm are possible but unlikely. Questions on gang membership have similarly implausible numbers, with about 6% of respondents reporting 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

[7]Folk wisdom gives "the mom who has to attend to the crying baby" and "the bored office worker whose boss dropped in unexpectedly" as potential reasons for long completion times.

[8]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).

"almost always" provided humorous or insincere responses to survey questions.

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

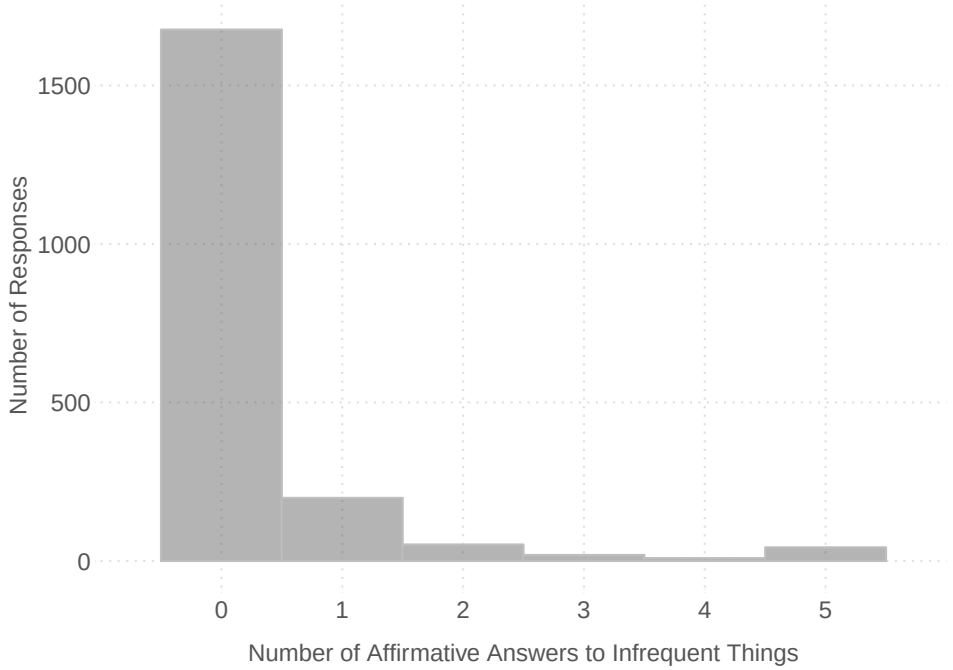
Rare Behavior/Trait	n
Use a Prosthetic	91
Blind	184
Deaf	109
Gang Member	88
Family Member in Gang	123
Sleep 10+ hrs or <4 hrs	28
2 or more of above	125

When we look at the association between admitting to providing insincere responses and responding affirmatively to questions about rare afflictions, traits, and behaviors, a pattern emerges. Of the 1,875 respondents who responded affirmatively on one or fewer on questions about rare traits and afflictions, nearly 93% reported that they "never" or "rarely" answered humorously or insincerely. By contrast, of the 125 who gave an affirmative answer to two or more such questions, roughly 58% said that they usually answered sincerely ($\chi^2 = 179.0, p < 0.001$). In all, we suspect that 5–7% of the `Workers` recruited for this study engaged in trolling or satisficing.

To assess covariation between various indicators of low-quality responding, we cross-tabulated the data on IP addresses with responses to questions about rare behaviors, traits, and conditions. Thirty eight (38) of the 408 responses from "bad" IPs (or about 9% of the sample) replied in the affirmative on two or more of our low incidence screener questions. In comparison, nearly 6% of the responses from other IPs did the same. The difference is statistically significant but not eye-catchingly large. But neither did we expect it to be: 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.

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

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



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 the next section will show, these fast respondents 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. In sum, nearly a quarter of responses are potentially untrustworthy.

4 Consequences of Low Quality Responses

The results above suggest that there are at least three significant concerns with the survey data being collected on MTurk. First, a sizeable proportion of respondents took the survey from a location outside the United States. If—as we suspect—the majority of these

respondents are foreigners, many of our responses originate from outside the sampling frame. Second, a significant number of individuals filed multiple responses. Finally, a non-trivial proportion of people appear to respond in intentionally humorous ways.

The consequences of the first concern will vary depending on the survey. In many cases, we expect it to simply add noise to estimates. With regard to the sectond concern, if people are filing multiple responses from the same IP address, we expect (artificially) narrower standard errors.

But potentially graver consequences exist. To the extent that `Workers` respond "randomly" by rushing through the survey—or, for foreign `Workers`, because they do not understand English or American politics—they introduce noise into the data. The noise will attenuate correlations and introduce bias in 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* error into estimates of the prevalence of certain attitudes (e.g., Lopez and Hillygus 2018). Depending on the relationship between the bias and the variables of interest, estimates can either be inflated or deflated. In either case, these errors threaten our ability to draw accurate inferences.

To get a better sense of how low-quality responses may influence the substantive conclusions reached in a study, we embedded an experiment on partisan stereotyping 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) examine 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 the demographic compositions of the Republican and Democratic parties. Specifically, they examine how likely respondents

are to 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 occurring separately (Tversky and Kahneman 1974), after being exposed to party-representative characteristics (like gender, race, sexual orientation, and religion).

Ahler and Sood (2017) introduced 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 the vignette, Ahler and Sood (2017) 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 (Democratic)

Party will always be less than or equal to the probability that he is either a salesman or a member of the Republican (Democratic) 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 estimate the impact of low-quality responses, 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 to analyze the data.[9] 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

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

operationalize `low quality responses` three ways in three different models: first as all respondents flagged for any reason, then as duplicated/flagged IP addresses, and finally as respondents flagged for potential trolling.[10]

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 the predicted probability of committing the Democratic/Republican conjunction fallacy. We first present results for all non-flagged respondents (column 1) and then by all low-quality respondents (which include flagged IP addresses and respondents we suspect are responding non-seriously (column 2). Finally, we present the results among 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 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 suspicious and non-suspicious respondents is notable. Suspicious respondents, for example, are nearly eight percentage points less likely

[10]In a fourth model, presented in SI 1.4, we assess the effect of speeding through the survey—operationalized as being a "fast outlier," or taking roughly four minutes or less. Unlike the effects shown in Table 3, fast responses do not appear to attenuate treatment effects.

Table 3: Impact of Low-Quality Responding on Treatment Effects - Marginal Effects

	Non-flagged respondents ($n = 1,446$)		All low-quality 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 ($p < 0.1$).

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 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. Averaging these differences in treatment effects, weighted inversely by their estimated standard errors, yields a precision-weighted average difference of 4.0 percentage points in average treatment effects between suspicious and non-suspicious respondents (95% confidence interval (CI): [0.4, 7.6]). 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: [0.1, 1.7]). We can contextualize this attenuation effect by putting it in percentage terms: the observed precision-weighted average treatment effect among non-suspicious respondents is 8.8 percentage points, and the presence of suspicious respondents (and their noisy data) attenuates this estimated effect by 8.6% (see SI 1.3 for more on this estimation procedure).

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 potential non-serious respondents. While the results among 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.

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. Seventy four percent (74%) of respondents committed one form of the conjunction fallacy in the James experiment. Fast outliers were 3.4 percentage points less likely to do so, but this apparent difference is imprecisely estimated (95% CI: [-0.17, 0.10]). 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. In only one out of six cases do they appear to respond significantly differently—the atheist/agnostic cue ($p = 0.09$)—but the coefficient is incorrectly signed for our hypothesis: fast outliers are slightly more responsive to this cue than slower non-suspicious respondents are.

One reason for this anomaly could be that people who complete surveys more quickly are better readers, and therefore comprehend vignette treatments in shorter amounts of time. We find that respondents who completed college—our best proxy for reading comprehension—are indeed 1.8 percentage points more likely to end up as 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.

5 Discussion and Conclusion

Our results suggest that cheating and trolling/satisficing are significant problems 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 other subjects. Specifically, we find that suspicious behavior—either in the form of cheating or trolling/satisficing—adds noise to the data, which attenuates treatment effects—in our case, by nearly 9%.

Current data quality may be low, but what's the prognosis? Given strategic incentives, we forecast steadily declining `Worker` quality on MTurk. 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 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 incentivize workers by giving them a bonus when their responses pass quality filters. That is, researchers could 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 unusually 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 and become respondents 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 filter to 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.

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 significant non-serious or fraudulent responding is a concern.

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 Gaurav Sood. 2017. Typecast: Cognitive Roots of Party Stereotyping. 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.

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.

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.

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).

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.

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.

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'.". 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.

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 Low Incidence Screener (Trolling) Question Text

- Do you use an artificial limb or prosthetic?—Yes, No
- Are you blind or do you have vision impairment?—Yes, No
- Are you deaf or do you have hearing impairment?—Yes, No
- Are you in a gang?—Yes, No
- Is one or more of your immediate family members in a gang?—Yes, No
- Finally, we sometimes find people don't always take surveys seriously, instead 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	-0.32	-0.36	-0.24
	(0.26)	(0.30)	(0.59)
Black	-0.60	-0.59	-0.59
	(0.10)	(0.10)	(0.10)
Black * LQ	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 * LQ	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. * LQ	-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 * LQ	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. * LQ	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. * LQ	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: "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	-0.60***
	(0.10)
Black * fast	0.96
	(0.97)
Gay	-0.80***
	(0.10)
Gay * fast	-0.15
	(0.86)
Evangelical	0.26**
	(0.12)
Evang. * fast	-0.52
	(1.00)
Atheist/agnostic	-0.27
	(0.13)
AA * fast	-1.83*
	(1.08)
Liberal	-0.44***
	(0.13)
Lib. * fast	-0.52
	(1.06)
Conservative	0.35
	(0.13)
Con. * fast	-0.96
	(0.96)
Cut 1	-0.58
	(0.14)
Cut 2	0.65
	(0.14)
Pseudo R^2	0.05
n	1482