Political Science Research and Methods (2021), page 1 of 20
doi:10.1017/psrm.2021.57

ORIGINAL ARTICLE

The micro-task market for lemons: data quality on Amazon's Mechanical Turk

Douglas J. Ahler[1]*, Carolyn E. Roush[1] and Gaurav Sood[2]

[1]Florida State University, Tallahassee, FL, USA and [2]Independent Researcher
*Corresponding author. Email: dahler@fsu.edu

(Received 22 January 2020; revised 29 July 2020; accepted 29 September 2020)

Abstract
While Amazon's Mechanical Turk (MTurk) has reduced the cost of collecting original data, in 2018, researchers noted the potential existence of a large number of bad actors on the platform. To evaluate data quality on MTurk, we fielded three surveys between 2018 and 2020. While we find no evidence of a "bot epidemic," significant portions of the data—between 25 and 35 percent—are of dubious quality. While the number of IP addresses that completed the survey multiple times or circumvented location requirements fell almost 50 percent over time, suspicious IP addresses are more prevalent on MTurk than on other platforms. Furthermore, many respondents appear to respond humorously or insincerely, and this behavior increased over 200 percent from 2018 to 2020. Importantly, these low-quality responses attenuate observed treatment effects by magnitudes ranging from approximately 10 to 30 percent.

Keywords: Experimental research; public opinion

1 Introduction

Over the past decade, Amazon's Mechanical Turk (MTurk) has dramatically changed social science. The platform has freed researchers from reliance on the "narrow database" of social science undergraduates (Sears, 1986) while reducing the cost and inconvenience of gathering original data (e.g., Berinsky *et al.*, 2012; Casler *et al.*, 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., Mullinix *et al.*, 2015; Hauser and Schwarz, 2016; Thomas and Clifford, 2015) and exhibit the same cognitive biases as participants recruited through more traditional means (e.g., Paolacci *et al.*, 2010; Horton *et al.*, 2011; Goodman *et al.*, 2012). 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. Over the past few years, researchers have discovered that a non-trivial proportion of MTurk data 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 to survey questions, they could introduce noise that would bias treatment effect estimates 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—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 three original studies—one in August 2018 and two in the summer of 2020—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 from which the survey was completed, to more closely examine responses. We also used survey completion times to identify potential survey satisficers. Finally, we included a battery designed to indirectly assess how many Workers engaged in "trolling"—that is, provided humorous or insincere responses to survey questions.

In our first survey, we found that 11 percent of respondents circumvented location requirements or used multiple devices to take the survey from the same IP address, while 16 percent of responses came from blacklisted IP addresses. Approximately 6 percent of respondents also engaged in trolling or satisficing. In all, when researchers first observed the data quality problem, about 25 percent of responses collected on MTurk appeared untrustworthy, a noteworthy uptick compared to studies conducted on the platform in 2015. While the rate of responses coming from suspicious or duplicated IP addresses fell between 2018 and 2020, according to our two additional studies, it remains three to five times higher than one would find on the least costly online survey panels (e.g., Dynata, Lucid). Even more troubling, the apparent prevalence of trolling on MTurk has tripled over the past few years.

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 an experiment embedded in our June 2020 study. Specifically, they attenuate treatment effects—in our case by 28 percent—by introducing noise into the data. This suggests that researchers' statistical power to detect effects is likely lower than implied by the observed *n*.

While we find relatively low response quality, researchers can preempt bad actors, most notably by restricting MTurk surveys to Workers with a long history of participation on MTurk. But tradeoffs exist: while data quality appears to improve significantly when we restrict surveys to Workers who have completed more than 1000 Human Intelligence Tasks (HITs), limiting participation in this way risks obtaining a sample comprised of Workers who may be overly familiar with surveys and who may be more subject to demand effects. Furthermore, the cost of conducting survey research on higher quality, centrally managed alternatives appears to be decreasing. As we show, samples recruited through Lucid (e.g., Coppock and McClellan, 2019; Graham, 2021, 2020; Thompson and Busby, 2020) cost roughly the same per valid response than those recruited through MTurk. As such, we recommend that researchers employ a broader range of databases when recruiting respondents and use MTurk thoughtfully, primarily for quick tests and pilots. We conclude by offering a number of recommendations for maximizing data quality in these contexts.

2 Incentives for quality on MTurk

MTurk is a micro-task market: people complete 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

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

the location from which they are taking the survey, unlike survey sampling firms who recruit "panelists" based on such prior information. Workers plausibly exploit this opacity for gain. For example, foreign nationals might 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 repeatedly, 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 could not be automated (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 lack an external, objectively correct answer. This makes it difficult to parse genuine responses from insincere or inattentive 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 et al., 1996; Vannette and Krosnick, 2014).

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, which has two significant consequences. First, professionally managed survey platforms recruit respondents based on known characteristics, which naturally culls respondents who misrepresent themselves. Second, if the typical researcher uses MTurk two to three times a year, they have few incentives to sink resources into monitoring quality; instead, their investment is typically capped at the payout rate. On the other hand, survey vendors' business model is based upon consistently providing high-quality data to clients. Since respondents take many surveys, these firms have opportunities to aggregate what might otherwise be individual weak signals into a more complete profile of respondent behavior which they can monitor. Participant monitoring that is not possible on MTurk—from recruitment through panel management—may incentivize respondents to behave more honestly.

When it comes to MTurk, the only signal of Worker quality that Requesters can send to the market is HIT approval—whether or not the Worker completed the task as assigned, which Amazon aggregates and tracks for each Worker. 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 limit this metric's usefulness for social science research.

Worse, HIT completion rates themselves are likely upwardly biased, weakening any potential signal they send. Not only is spot-checking data for response quality time consuming for Requesters, but flagging false positives can hurt Requesters' reputations and hence impose future costs. Workers who are denied a payment 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 a few cents—Requesters often batch approve completed HITs, making the HIT completion metric a biased signal of Worker quality.

This information asymmetry gives Workers incentives to misrepresent where they are located, use multiple accounts 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—that is, providing outrageous or misleading responses—as they can make

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

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

money while indulging their id (e.g., Cornell *et al.*, 2012; Robinson-Cimpian, 2014; Savin-Williams and Joyner, 2014; Lopez and Hillygus, 2018).

All of this suggests that data collected on MTurk may not be of the quality researchers often assume.[2] 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, bad actors may be far more common than is typically assumed. Moreover, these incentives may lead to further deterioration of data quality on the platform over time. Just as "bad money drives out good," Gresham's law suggests that dishonest behavior from Workers may become the norm on MTurk, as there are few incentives for Workers to respond sincerely when satisficing, trolling, speeding through surveys, and circumventing location requirements saves them time and increases their profits.

Regardless of future trends, low-quality responding presents serious problems for social scientists using MTurk data *now*. Failing to exclude low-quality responses from MTurk data provides misleading estimates of scale reliability and introduces spurious associations between measures (Chandler *et al.*, 2020). As we detail below, low-quality responses are also liable to introduce significant noise into the data, not only reducing statistical power but also attenuating experimental treatment effects. For these reasons, understanding just how common the problem of low-quality responding on MTurk is can greatly improve the quality of inferences that scientists draw from MTurk data.

3 Assessing the quality of responses on MTurk

After becoming aware of the potential "bot" problem, we posted a survey on MTurk on August 17, 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 60 cents for each completed HIT. In keeping with best practices (Peer *et al.*, 2014)—and, thus, consistent practices, for external validity—we restricted participation to MTurk Workers with a HIT completion rate of at least 95 percent.[3]

First, to assess how many Workers use 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. Second, to identify people who masquerade as someone else or provide misleading data regarding the location from which they are taking the survey, we exploited data on IP addresses. 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,

[2] We should note that this paper addresses one specific subset of concerns about respondents and data quality: respondents being deceitful in one way or another. There are other dimensions of data quality, like respondent attentiveness. On this metric, MTurk consistently outperforms other platforms (Thomas and Clifford, 2017). However, it is likely that this metric is upwardly biased because "professional" respondents, common on MTurk because of qualification criteria, know that they must respond carefully to these questions.

[3] We did not restrict participation based on the number of previous HITs completed, as some have recommended (e.g., Amazon Mechanical Turk, 2019*b*). This is a sampling choice with inherent tradeoffs. On the one hand, our 2020 studies suggest that Workers with 1000 or more HITs appear to provide more genuine responses. On the other, they raise concerns about "professional survey responding." Respondents who take lots of surveys may be more *or* less likely to satisfice and may be less politically interested than other respondents, but they may also learn from surveys, and potentially become attuned to research hypotheses (i.e., "panel conditioning"; Krosnick, 1991; Hillygus *et al.*, 2014). We chose not to impose restrictions based on HITs in our first survey for three reasons: (1) the field lacks a body of research to say anything for sure about the consequences of professional responding (Hillygus *et al.*, 2014); (2) we expect many researchers will shy away from such restrictions based on the above concerns; and (3) our primary interest with this first survey was to glean the full scope of the problem as the typical researcher would encounter it. As we note in later analyses and in our recommendations for future research, restrictions based on HIT counts do appear to curtail suspicious responses. At the very least, Requesters should implement a filter of at least 100 HITs, since all Workers with fewer completed HITs are assigned a 100 percent approval rating by Amazon (Litman, 2019).

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

2006), the largest, most trusted provider of geoIP data, to provide locations of IP addresses. Know Your IP also collects data on blacklisted IP addresses, which often appear on the same traffic anonymization services that people use to evade location filters.[4] Know Your IP pulls blacklist data from ipvoid.com, which collates data from 96 separate blacklists.[5]

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 or residents of the same college dorm), 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.[6]

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. The median completion time was 573 seconds—or 9 minutes and 33 seconds, 27 seconds under the 10 minute target we provided. We flagged respondents as outliers if they finished 167 percent outside the interquartile range (IQR) of completion times.

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 for more than 10 hours or fewer than 4 hours as unusual. In keeping with previous research, we flagged respondents as satisficing or trolling if they provided affirmative answers to two or more of these items (Lopez and Hillygus, 2018).[7] 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 2.)

3.1 Study 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 recorded the IP addresses of 1,991 responses.[8] (We consider the nine responses for which Qualtrics could not record the IP address as suspect.)

[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.

[5] Know Your IP requires some familiarity with Python. Similar packages exist for use in R and Stata—see Kennedy *et al.* (2020) for one example.

[6] A cursory look at the start and stop times on responses originating from the same IP address suggests many of these multiple submissions are being completed by the same person, potentially on multiple devices; in most instances, multiple submissions begin or end within a minute of one another. Even if these are the result of one Worker alerting another in the same location to our HIT, we would ideally take this into account when making standard error calculations. See SI 1.2 for a full accounting of these data.

[7] 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.

[8] The MaxMind algorithm could not reliably ascertain a latitude and longitude specific enough to pin provide a location for two IP addresses. This is because MaxMind collects information on IP addresses from multiple sources, which may not

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

Table 1. Frequency of different types of suspicious IPs, Study 1

Type of suspicious IP	Percentage of data
Missing	0.5%
Blacklisted	16.1%
Duplicated	5.3%
Foreign	6.0%
Any of the above	20.3%
n	2,000

Of the 1,991 responses, approximately 5 percent came from an IP that appears in our dataset more than once (see Table 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 complete the HIT multiple times.[9]

A large majority of responses (94 percent) originated from within the United States (see Table 1). Of the 125 foreign responses, roughly a third were from Venezuela and an additional 13.6 percent were from India. (See Table SI 1.2 for a complete distribution of countries from which HITs were completed.) We suspect that these 125 responses are from MTurk Worker accounts that were created using US 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 Kennedy *et al.* (2018) and Ryan (2018) (see Table SI 1.3). Yet more shockingly, of the responses with recorded IP addresses, 16 percent come from blacklisted IPs. In all, around 20 percent 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 the survey. We found that just over 2 percent of respondents were "fast outliers" who completed the survey in under 245 seconds. Consistent with folk wisdom, far more respondents (11.7 percent) were classified as "slow outliers."[10]

Next, we examined the frequency of insincere or inattentive respondents. Just over 9 percent of respondents in our data report being blind or having a visual impairment (see Table 2). Another 5.5 percent report being deaf. These numbers are nearly three and 14.5 times their respective rates in the population.[11] These large deviations from the national norm are possible but unlikely. Questions on gang membership have similarly implausible numbers (National Gang Intelligence Center (US) 2012). To be cautious, however, we only flag a respondent as trolling if they answered "yes" to two or more on such items. (See Figure SI 1.2 for the distribution of affirmative responses to these questions across all studies.) In all, we classify roughly 6 percent of respondents as likely "trolls."

Additionally, roughly 9 percent of respondents reported that they "always" or "almost always" provide 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 those who responded affirmatively to one or fewer low-incidence screeners, nearly 93 percent reported that they "never" or "rarely" answered humorously or insincerely. By contrast, roughly 58 percent of the 125 classified as trolls said that they usually

comport—especially if technology is being used to mask the user's true location. Accordingly, some analyses that follow are based on the full sample (all $n = 2000$ cases sampled), while others are based on all cases for which we captured an IP address ($n = 1991$) and still others are based on all cases for which we could reliably locate the respondent ($n = 1989$).

[9] All studies in this paper took under 3.5 hours to sample the desired *n*. Observing multiple submissions from the same IP address without coordination over this time window seems unlikely.

[10] 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, for example, by a crying baby or an uncomfortably proximate boss.

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

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

Table 2. Frequency of rare behaviors/traits, Study 1

Rare trait/behavior	Percentage of data
Prosthetic	4.6%
Blind	9.2%
Deaf	5.5%
Gang member	4.4%
Family is gang member	6.2%
Sleep < 4 or 10 + hours/night	1.4%
Two or more of the above	6.3%
n	2,000

answered sincerely ($\chi^2 = 166.2$, $p < 0.001$). In all, about 6 percent of Workers recruited for this study potentially responded insincerely.

To assess associations between measures of low-quality responding, we compare flagged IP addresses to Workers flagged as likely "trolls." Thirty eight (38) of the 406 responses from "bad" IP addresses (about 9 percent of the sample) replied affirmatively to two or more of these items, compared to a rate of 5.5 percent among non-suspicious IP addresses. This difference is statistically significant ($p < 0.05$) but not immense. 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, however, is another question.

Surprisingly, we find that, on average, potential trolls and responses from questionable IP addresses take significantly longer to finish (by 166 seconds, $p < 0.001$) and are significantly more likely to be slow outliers ($\hat{\beta} = 0.14$, $p < 0.001$). On the other hand, they are no less likely to be fast outliers ($\hat{\beta} = -0.00$, $p = 0.69$). We therefore do not count fast outliers as untrustworthy responses. And, as we show in SI 1.5, unlike other flagged respondents, speedsters do not appear to provide lower-quality data.

In all, about 25 percent of responses are from IPs that are duplicated, located in a foreign country, or blacklisted, or come from respondents who provided affirmative answers to two or more of the low-incidence questions. Altogether, nearly a quarter of responses are potentially untrustworthy.[12]

3.2 Results from studies 2 and 3

One hopeful possibility is that data quality on MTurk was uniquely bad in 2018. That is, the problem may have been detected when data quality fell noticeably, and the collective response (e.g., Amazon Mechanical Turk, 2019*a*, 2019*b*) restored data quality to previous levels. To revisit the question, we fielded two new surveys in the summer of 2020. In June, we paid respondents ($n = 1{,}503$) 75 cents to complete a 15-minute survey (Study 2), which included an experiment (detailed in Section 5), the noCAPTCHA reCAPTCHA qualification (again, all respondents passed), and the low-incidence screener battery, among other items. In July, we paid respondents ($n = 409$) 35 cents to complete a 5-minute survey (Study 3) with relatively little content aside from noCAPTCHA reCAPTCHA (again, all passed) and items to assess data quality. Importantly, to determine the efficacy of HIT approval rates for parsing bad actors from the rest, we restricted Study 3 to Workers with a 95 percent or higher HIT approval rate but did not do so for Study 2.

Table 3 demonstrates that the number of responses from suspicious IP addresses has changed substantially between 2018 and 2020. Using the most apt comparison—Study 3, which imposed

[12] Though this figure seems rather high, it may in fact underestimate 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 percent of participants passed a manipulation check (Appendix B). With few incentives for survey respondents to carefully read and process every question, many Workers recruited for our study may have also failed to attention to portions of our survey (although see Thomas and Clifford (2017), on MTurk respondents' apparent attentiveness).

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

Table 3. Frequency of different types of suspicious IPs

Survey	Missing	Blacklisted	Duplicated	Foreign	Any	*N*
August 2018 (Study 1)	0.5%	16.1%	5.3%	6%	20.3%	2,000
June 2020 (Study 2)	0%	6.7%	5.4%	1.1%	12.1%	1,505
July 2020 (Study 3)	0%	5.9%	3.2%	0.7%	9.8%	409

the same qualification restrictions as Study 1—we see a marked reduction in the proportions of responses originating from blacklisted, foreign, or duplicate IP addresses. In particular, the proportions of blacklisted and foreign IP addresses found in Study 3 fell by more than a third and by roughly 90 percent, respectively, compared to Study 1. In total, about 10 percent of the data in July 2020 originated from suspicious IP addresses compared to about 20 percent in August 2018. Even Study 2, which did not impose restrictions based on HIT approval rates, received substantially fewer responses from blacklisted, foreign, and duplicate IPs than the 2018 survey.[13]

But any reductions in responses originating from suspicious IP addresses are offset by increases in humorous or insincere responding. These changes over time are cataloged in Figure 1. Whereas approximately 6 percent of the 2018 sample reported having two or more uncommon traits or engaging in two or more suspect behaviors, 21 percent of survey takers did the same in July 2020 (and 30 percent of the unrestricted sampled did so in June). Just under 5 percent of respondents in 2018 claimed to use a prosthetic; this rose to 20 percent in 2020 (and to 27 percent with less stringent respondent restrictions). A similar proportion of respondents claimed gang membership, while just 6 percent did so in 2018. And, as further evidence that at least some of this uptick is due to trolling and not just inattentive responding, over 14 percent of both 2020 samples admitted to responding inaccurately and humorously on surveys, compared to 8.8 percent in 2018. So while there may be fewer foreign respondents to add noise to the data, there appear to be significantly more insincere respondents. Crucially, with greater potential to produce systematic error, trolls may do more damage to survey quality than those who respond randomly.

Additionally, evidence from the 2020 surveys suggests that there may be more foreign respondents than we estimate based on IP address data alone. Namely, both surveys included a new item designed to detect responses that potentially originate outside the US. One unique US convention is how the date is written: nearly all other countries use DD/MM/YYYY as a shorthand format, while the U.S. uses MM/DD/YYYY. We asked respondents the following:

Please write today's date in the text box below. Be sure to type it using the following format: 00/00/0000.

Approximately 17 percent of responses to this item in our July survey (Study 3) were written in the format DD/MM/YYYY, a number that rises to 20 percent in our June survey (Study 2) without the HIT approval rate qualification. In addition to dates that were written in the DD/MM/YYYY, we found that an additional 4.2 percent of respondents in Study 3 and 4.9 percent of respondents in Study 2 did not write anything resembling a date in the allotted space. This suggests that these are respondents who are taking the survey inattentively.

What explains the high proportions of survey takers who wrote the date in a format uncommonly used in the U.S.? Two major possibilities exist. One is that foreign respondents are able to use a VPN service to circumvent Amazon's location filter, appearing to have taken the survey in the U.S. when it was taken from a location outside the country. Another is that a disproportionate number of MTurk Workers in the U.S. are immigrants who retain their original date-writing custom.[14]

[13] For more about fast and slow outliers in these studies, their similarities to timing outliers in 2018, and their reactions to experimental treatments compared to other respondents to the June 2020 survey, see SI 1.5.

[14] Interestingly, people who wrote the date this way took marginally longer, on average, than those who used the American convention. This suggests some may have attempted to read the survey and answer in a systematic fashion. See SI 1.3 for a complete accounting.

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

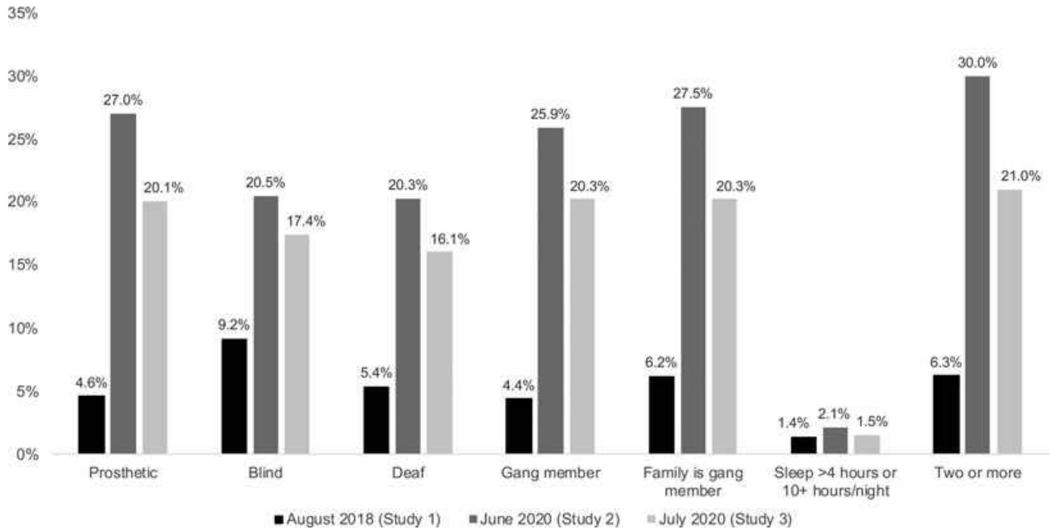


Figure 1. Frequency of rare behaviors/traits across three studies. *Note*: Studies 1 and 3 imposed a 95 percent HIT completion rate restriction; Study 2 did not.

Table 4. Estimates of low-quality responding using various thresholds

	August 2018 (Study 1)	June 2020 (Study 2)	July 2020 (Study 3)
% Low-quality, index 1	24.7%	37.9%	27.6%
% Low-quality, index 2	NA	44.6%	35.2%
% Low-quality, index 3	NA	46.3%	35.2%
n	2,000	1,505	409

Note: Studies 1 and 3 imposed a 95% HIT completion rate restriction; Study 2 did not.
% Low-quality, index 1 includes suspicious IPs and incidences of trolling (per low-incidence screener measures); *% Low-quality, index 2* includes suspicious IPs, incidences of trolling, and incidences of the date written in the DD/MM/YYYY format; *% Low-quality, index 3* includes suspicious IPs, incidences of trolling, incidences of the date written in DD/MM/YYYY format, and any non-date entered into the open-ended date question.

Unfortunately for scholars of American politics, we cannot parse these possibilities. That being said, this survey item does allow us to put an "upper bound" on our estimates of undesirable responses. Table 4 provides more clarity about the overall frequency of low-quality responses in our data using different metrics based on the information described above. The "lower bound" of low-quality responses, classified as *% Low-quality, index 1* in the table, is based on our original definition of "low quality": it includes the proportion of responses that either originated from suspicious IP addresses or were flagged as potential trolls by having answered in the affirmative to two or more low-incidence screener questions. *% Low-quality, index 2* includes the former two qualifications but also adds in any respondents who wrote a date in the DD/MM/YYYY format. Finally, *% Low-quality, index 3* gives us an "upper bound" on our estimate of low-quality responses by including aforementioned qualifications and any additional responses that did not provide a date when asked to do so.

These results suggest that large proportions of data collected on MTurk are of low-quality. About 25 percent of responses in Study 1 (conducted in August 2018) originated from suspicious IP addresses or were flagged as potential trolls; two years later (in Study 3), the proportion of responses flagged according to the same criteria was about 27 percent. While this may not be much of an increase, as discussed previously, these data belie the fact that the number of suspicious IP addresses on MTurk has decreased while estimates of trolling have increased by nearly fourfold (see Figure 1). When including other responses that used the DD/MM/YYYY format or

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

Table 5. Low-quality responses by HIT completion rates

	Fewer than 100 HITs	Between 100 and 500 HITs	Between 500 and 1k HITs	More than 1k HITS
June 2020 (Study 2)				
% of sample	21.8%	29.0%	12.8%	36.5%
% Low-quality, index 1	56.0%	51.5%	45.3%	14.1%
% Low-quality, index 2	68.8%	60.5%	51.6%	15.3%
% Low-quality, index 3	70.0%	60.9%	53.7%	17.5%
% Self-reported insincere	14.7%	23.0%	16.7%	6.6%
Average completion time	678.3	608.5	586.2	495.9
July 2020 (Study 3)				
% of sample	20.7%	26.9%	9.4%	43.1%
% Low-quality, index 1	50.0%	40.4%	31.6%	6.9%
% Low-quality, index 2	63.1%	51.4%	34.2%	10.9%
% Low-quality, index 3	63.1%	51.4%	34.2%	10.9%
% Self-reported insincere	14.3%	24.8%	34.2%	3.4%
Average completion time	544.9	490.8	452.2	280.5

Note: Studies 1 and 3 imposed a 95% HIT completion rate restriction; Study 2 did not.
% Low-quality, index 1 includes suspicious IPs and incidences of trolling (per low-incidence screener measures); *% Low-quality, index 2* includes suspicious IPs, incidences of trolling, and incidences of the date written in the DD/MM/YYYY format; *% Low-quality, index 3* includes suspicious IPs, incidences of trolling, incidences of the date written in DD/MM/YYYY format, any non-date entered into the open-ended date question.

answered our date question nonsensically, our estimate of low-quality responses increases to roughly 35 percent.[15] Both of these studies required Workers to have at least a 95 percent HIT completion rate, and this qualification requirement makes a difference in the proportion of low-quality responses. Without the qualification requirement, roughly 38 percent of the data in Study 2 originated from suspicious IP addresses or from potential trolls. When we add in those respondents who did not answer the date question according to directions (or in the conventional U.S. format), our estimate of low-quality responses jumps to nearly 50 percent. In sum, the proportion of low-quality data on MTurk ranges between 25and 50 percent, depending on the qualification requirements (based on HIT approval rate) imposed.

While these figures are troubling, perhaps the prevalence of low-quality responding can be reduced by requiring Workers to complete at least 5,000 HITs, as Amazon now recommends (Amazon Mechanical Turk, 2019*b*). In our 2020 surveys, we asked respondents to review their MTurk Worker account page and report the rough number of HITs they had completed.[16] In Table 5, we parse low-quality responses by the reported number of completed HITs. Indeed, data quality appears significantly better among Workers with at least 1,000 completed HITs. Our "upper bound" of bad actors topped out at roughly 10 percent in this subgroup in Study 3—not ideal, but far better than the roughly 35 percent that we got in the full sample. Here, the 95 percent HIT approval rating appears to make a difference as well: our "upper bound" estimate of low-quality responses increased to nearly 18 percent in Study 2, which did not impose the 95 percent approval rate restriction. And, when it comes to trolling, just 3.4 percent of those Workers who completed more than 1,000 HITs admit to responding insincerely when asked. (Without the 95 percent completion rate restriction, this figure rises to 6.6 percent.)

Consistent with Zhang *et al.* (2020), then, we find that respondents who have taken lots of surveys tend to provide higher-quality data by observable measures. There is reason for caution, however. Respondents who reported having completed more than 1,000 HITs finished our surveys

[15] All responses in Study 3 that included a non-date answer to the open-ended date question were already classified as low-quality by having originated from a suspicious IP, by answering in the affirmative to two or more low-incidence screener questions, or by having written the date in the DD/MM/YYYY format.

[16] See SI 2.3 for question wording.

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

significantly more quickly than those who have completed fewer HITs. On average, these individuals took about 280 seconds to complete Study 3, making them 38 percent faster than those who reported completing 500–1,000 HITs. In Study 2, these high-HIT Workers were 15 percent faster than the 500–1,000 HIT respondents. While we cannot be certain, the fact that Workers with large numbers of completed HITs complete these surveys more quickly than other respondents suggests that these Workers may be gaining expertise in survey-taking if they complete enough HITs. These individuals may become better attuned to research hypotheses, which presents its own threats to validity (e.g., Campbell and Stanley, 1963). Researchers should be mindful of these tradeoffs when designing studies on MTurk.

4 A market for lemons?

Thus far, we have provided some cursory evidence that low-quality responding on MTurk has increased over time. We have also theorized that low-quality responding may be more prevalent on MTurk than on other, more centrally managed survey platforms due to its unique incentive structure and information asymmetries. Now, we bring more evidence to bear on these questions by comparing rates of low-quality responses in MTurk studies to rates of low-quality responses on surveys conducted on Dynata (formerly Survey Sampling International) and Lucid, two relatively low-cost online panels.

To do so, we make use of data generously provided by other researchers. The first three datasets come from studies conducted in 2015. Two were administered using MTurk (Ahler and Broockman, 2018; Ahler and Goggin, 2019) and the other using SSI/Dynata (Institute of Governmental Studies at the University of California, Berkeley, 2015). The other four studies—Busby (2020), Graham (2021), Graham (2020), and Thompson and Busby (2020)—were administered by Lucid either in 2019 or 2020.[17] While these studies did not include a trolling battery, we are able to provide an estimate of low-quality data based on suspicious IP addresses. As before, we made use of Know Your IP (Laohaprapanon and Sood, 2018) to determine the proportion of respondents in each study who circumvented location requirements, completed the survey more than once, took the survey from a location outside the US, or completed the survey using a blacklisted IP address.

Is the problem of responses from suspicious IP addresses unique to MTurk? The data presented in Table 6 suggest this is the case. Looking back, we can see that the prevalence of suspicious IP addresses on MTurk did not suddenly increase in 2018. Roughly 18 percent of the data in both Ahler and Broockman (2018) and Ahler and Goggin (2019) appear to originate from suspect IP addresses. This is not a far cry from the rate of bad IP addresses that we identified in Study 1 in 2018 (roughly 20 percent). And our estimates for the 2015 studies likely underestimate 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. By contrast, only about 4 percent of the data collected by SSI/Dynata for the Berkeley IGS Poll were flagged for suspicious IP addresses.

As noted previously, the prevalence of suspicious IP addresses on MTurk has decreased significantly since 2019, when Amazon implemented restrictions to weed out bad actors (Amazon Mechanical Turk, 2019*a*, 2019*b*); the proportion of responses originating from suspicious IP addresses decreased by 50 percent among Workers with a 95 percent HIT approval rating between Study 1 and Study 3. That being said, MTurk still appears to disproportionately elicit

[17] It’s worth noting that we explore only certain types of low-quality responding here. Outside the scope of our investigation here is inattentiveness, and there are recent concerns about that particular problem on Lucid (Aronow *et al.*, 2020). Inattentiveness theoretically introduces similar error into survey data as foreign respondents who do not understand the questions, and contemporary data on various samples’ attentiveness should be considered when devising sampling strategies and conducting power analyses.

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

Table 6. Frequency of suspicious IP addresses across different platforms

Survey	Year	Platform	% Suspicious IP addresses
Berkeley IGS Poll	2015	SSI (Dynata)	4.3%
Ahler and Broockman	2015	MTurk	17.7%
Ahler and Goggin	2015	MTurk	17.9%
August 2018 Survey (Study 1)	2018	MTurk	20.3%
June 2020 Survey (Study 2)	2020	MTurk	12.1%
July 2020 Survey (Study 3)	2020	MTurk	9.8%
Busby	2019	Lucid	2.2%
Graham	2019	Lucid	0.6%
Thompson and Busby	2019	Lucid	1.6%
Graham	2020	Lucid	1.5%

respondents with suspicious IP addresses compared to Lucid, another relatively inexpensive survey platform. We estimate that there are still over five times as many bad actors in MTurk samples than in Lucid samples in 2020—and this is based only on suspicious IP addresses. While we suspect some Lucid survey-takers also engage in trolling or satisficing, the fact that Lucid institutes periodic quality checks on panel participants may incentivize respondents to behave more honestly.

Many researchers have turned to MTurk because it provides a low-cost alternative to curated, representative samples maintained by large survey firms. Unfortunately, it may no longer be significantly more cost-effective than higher-quality alternatives, especially after accounting for the 30 percent surcharge Amazon imposes on Requesters. Our June 2020 study paid respondents 75 cents per completed survey; the cost increased to about 98 cents per respondent after accounting for the surcharge. Our estimates suggest that 12 percent of Workers who completed that survey originated from suspicious IP addresses. In order to obtain a cost-per-minute for a valid survey, we divided the cost per non-suspicious response (97.5 cents) by the proportion of responses that were non-suspicious (0.879), and then divide that quantity by the survey length (15 minutes, both advertised and in reality, on average)—obtaining a cost of about 7 cents per minute. Following the same procedure for our July 2020 survey, we obtain a cost of $\frac{0.35*1.3}{0.902} \div 5\text{minutes} = 10$ cents per minute. By contrast, making use of information from Busby (2020), Thompson and Busby (2020), Graham (2021), and Graham (2020), we estimate that researchers can obtain samples from an apparently higher quality panel at a similar cost: with a cost of $1 per respondent, just 2 percent of responses coming from suspicious IP addresses (see Table 6), and both surveys clocking roughly 10 minutes, we estimate a cost of approximately 10 cents per valid response per minute on Lucid.

This may, however, be a low estimate. As Thompson and Busby (2020) describe, Lucid can be more expensive when sampling specific subgroups—that study paid $2 per respondent to sample only white Americans. At 10 minutes and with 1.6 percent of the sample coming from non-suspicious IP addresses, that study appeared to pay just over 20 cents per valid response. But the valid comparison on MTurk may involve greater costs as well, with more unknown unknowns: if a Requester attempts to restrict her HIT to a particular subgroup, she is liable to sample a significant number of people masquerading as members of that subgroup.

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, even with location filters and improvements to the system since 2019, a non-trivial number of MTurk respondents take surveys from outside the United States. If—as we suspect—the majority of these respondents are foreign, many of our responses are provided from people from outside the sampling frame. Second, many respondents filed multiple

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

submissions. Finally, a significant and rising proportion of MTurk Workers appear to provide intentionally humorous or misleading answers to survey items.

Some may assert that these problems are mere annoyances, as random responding could simply add "noise" to the data. But this noise itself may be a bigger problem than many assume. Not only can this imprecision bias estimates of frequencies and means of some measures—for instance, even answering questions randomly can positively bias estimates of how many people know something (Cor and Sood, 2016)—but it can also attenuate correlations. In an experimental context in which researchers have control over the independent variable, T_i—where i indexes survey respondents, each randomly assigned to a treatment or control group—and only the dependent variable, Y_i, is vulnerable to noise, low-quality respondents will bias average treatment effects toward zero.[18]

Consider an experimental data-generating process for which $\beta_1 \neq 0$: there is an average treatment effect (ATE) that is $E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \neq 0$. When we randomly assign subjects, under usual conditions, we obtain $\bar{Y}_i|T_i = 0$ and $\bar{Y}_i|T_i = 1$, which we can use to compute an unbiased estimate of the ATE. But when there is haphazard responding by some subset of respondents j, $\bar{Y}_j$ is centered around neither $E[Y_i|T_i = 1]$ nor $E[Y_i|T_i = 0]$. And since $|E[\bar{Y}_i|T_i = 1] - E[\bar{Y}_i|T_i = 1]| > 0$ and $|E[\bar{Y}_j|T_i = 1] - E[\bar{Y}_j|T_i = 1]| = 0$, the average of the two will necessarily be smaller in absolute value than the former alone.

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 (e.g., Lopez and Hillygus, 2018). That is, in these cases, we might expect that a subset of respondents with a predilection for trolling, indexed by k, would provide average responses $E[Y_k|T_k = 1] \,!= E[Y_i|T_i = 1]$ and $E[Y_k|T_k = 0] \,!= E[Y_i|T_i = 0]$. Thus, trolling's effects are likely idiosyncratic to samples, treatments, and dependent measures. But either pitfall—attenuation of treatment effects from inattentive responding or biased effect estimates from trolling—threatens our ability to draw accurate inferences from an experimental design.

To study how low-quality responses influence the substantive conclusions reached in a study, we made use of data collected as part of an experiment on partisan motivated reasoning fielded by Roush and Sood (2020) in Study 2.[19] The experiment tested the hypothesis that partisans interpret the same economic data differently depending on the party to which economic success or failure is attributed. Specifically, Roush and Sood (2020) provided respondents with real economic data from 2016—collected by the Federal Reserve—demonstrating a decrease in the unemployment rate from 5.0 to 4.8 percent and a decrease in the inflation rate from 2.1 to 1.9 percent. Importantly, respondents were randomly assigned to receive one of two partisan cues preceding the question preamble: one-half of respondents were told that "During 2016, when Republicans controlled both Houses of Congress, [unemployment/inflation] decreased from X percent to X percent…" and the other half were told that "During 2016, when Barack Obama was President, [unemployment/inflation] decreased from Y percent to Y percent…" Respondents were then asked to interpret these changes and evaluate whether unemployment or inflation "got better," "stayed about the same," or "got worse." Since prior research demonstrates that partisans interpret economic conditions favorably when their own party is in power and unfavorably when the other party is in power (e.g., Bartels, 2002; Bisgaard, 2015), we expected that Democrats will be more likely than Republicans to interpret these slight decreases as improvements when they receive the President Obama cue, while Republicans will be more likely to interpret these same statistics positively when they receive the Republicans-in-Congress cue. In other words, we expected that partisans are more likely to classify a 0.2-point reduction in unemployment or inflation as "getting better" under a co-party president while interpreting this

[18] In observational studies, noise in the dependent variable "only" yields a larger variance of $\hat{\beta}$, while noise in the independent variable biases $\hat{\beta}$ toward zero, a phenomenon known as attenuation bias. As we demonstrate below, noise in the dependent variable can bias $\hat{\beta}$ toward zero in experiments as well.

[19] See SI 2.4 for full question wording.

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

same reduction as unemployment and inflation "staying the same" or "getting worse" under out-party leadership.

We recoded the data so that treatments and respondents are characterized in relation to one other: Democrats who saw the President Obama cue and Republicans who saw the Republicans-in-Congress cue were classified as receiving an *In-party cue*, whereas Democrats who saw the Republicans-in-Congress cue and Republicans who saw the President Obama cue are classified as receiving an *Out-party cue*. To determine how low-quality responses moderate this expected treatment effect, we adopted the following model:

$$\begin{aligned}\text{Economic Evaluation}_{\text{ij}} = \beta_0 + \beta_1 \text{In} - \text{party cue}_{\text{i}} + \beta_2 \text{LQ}_{\text{i}} \\ + \beta_3 (\text{In} - \text{party cue}_{\text{i}} \text{XLQ}_{\text{i}}) + \epsilon_{\text{ij}}\end{aligned} \tag{1}$$

where i indexes individual survey respondents, j indexes survey items (e.g., whether the respondent is assessing unemployment or inflation), and LQ_i is an indicator for low-quality response. We operationalized low quality four ways in four different models. The first includes respondents flagged for any reason; the second includes only those respondents who completed the survey from a suspicious IP address; the third includes only those respondents flagged for potential trolling; and the fourth includes only those respondents who self-reported completing 1000 HITs or more. All variables were recoded 0–1 for ease of interpretation.

5.1 Results

As Figure 2 shows, MTurk respondents who misrepresent themselves or answer questions insincerely can significantly affect experimental inference.[20] The first panel presents the results among the full sample, and the subsequent panels show the results for the other aforementioned subgroups. The second panel presents the experimental results among respondents not flagged for any reason. As expected, non-suspicious respondents who identified as Democratic or Republican—or as Independents who "lean" toward one of the parties—provided systematically different assessments of US economic performance depending on the partisan cue they received. These respondents evaluated the slight decrease in unemployment 12.5 percentage points more positively, on average, when the change was attributed to their own party (95 percent CI: [8.5, 16.4]). Similarly, they evaluated the marginal decline in inflation 9.1 percentage points more positively when presented with a cue implicitly crediting their own party (95 percent CI: [4.9, 13.2]).

The analogous treatment effects are decidedly smaller among the 579 respondents marked as suspicious. Panel 3 in Figure 2 presents these effects. Suspicious respondents responded to the treatment, albeit less strongly than non-suspicious respondents: partisans evaluated the 0.2 percentage point decrease in unemployment 5.3 points more positively when responsibility was attributed to their own party (95 percent CI: [0.8, 9.8])—a treatment effect consistent with the study's hypotheses, but also one significantly smaller than that among all non-suspicious respondents. We found no effect of the partisan cue on these respondents' evaluations of the change in inflation, however; the coefficient is neither statistically significant nor substantively meaningful.

The overall effect of suspicious respondents in the sample is an attenuation effect of 28 percent. The average treatment effect (ATE) of the out-party cue on assessments of the unemployment rate, among all respondents, is -0.10 (95 percent CI: $[-0.13, -0.07]$). The ATE of the out-party cue on assessments of inflation, again among the full sample, is -0.06 (95 percent CI: $[-0.09, -0.02]$). Among all suspicious respondents, the analogous ATEs are -0.05 and -0.02, as shown in panel 3 in Figure 2. We divide the estimated ATE among non-suspicious respondents by the estimated ATE in full sample to obtain the relative size of the observed effect to the "real" effect—the attenuation ratio. We calculate the average attenuation ratio across

[20] For full model results, see Table SI 3.5.

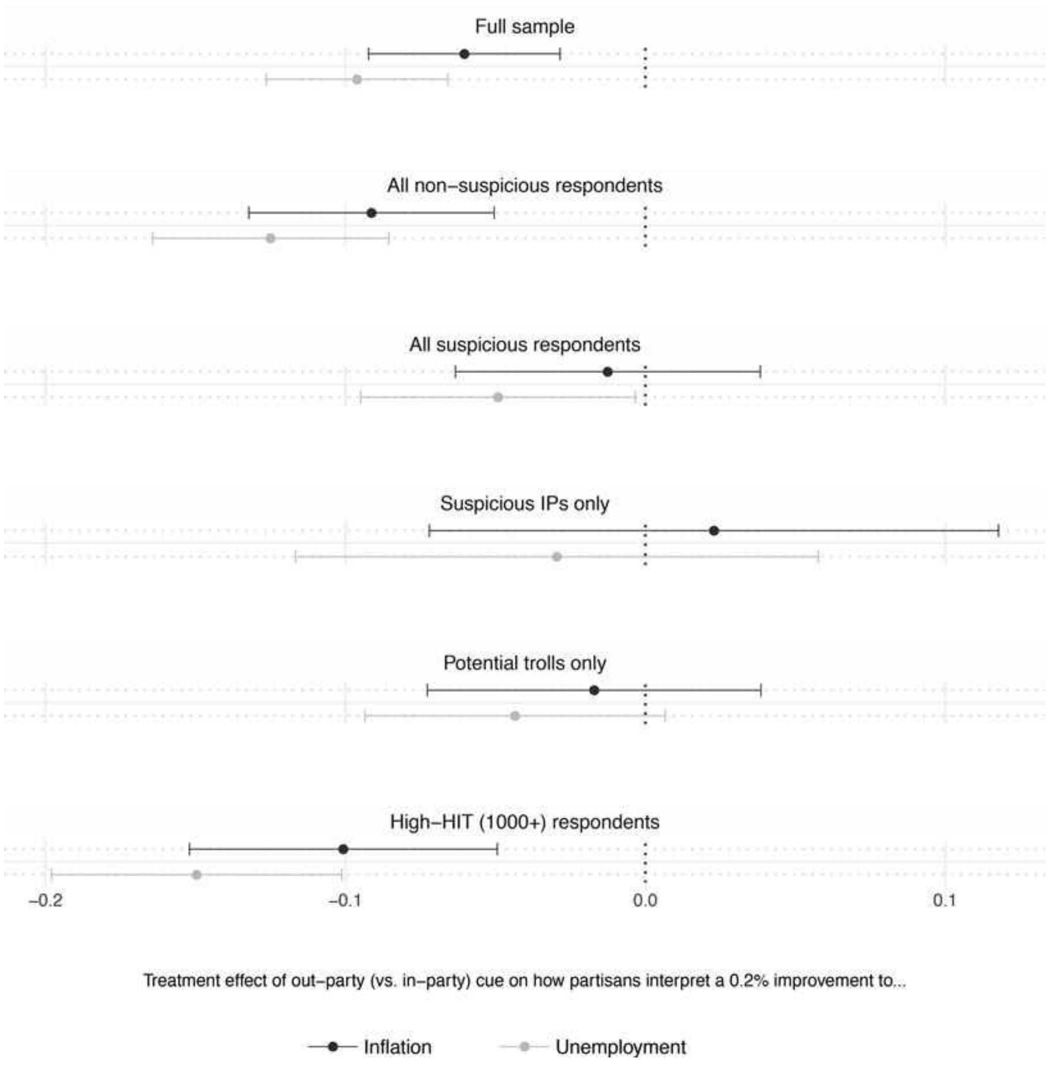


Figure 2. Experimental effects by subgroup.

outcome variables, weighted by the inverse of the standard error of these estimated ATEs, as 0.72; that is, our observed effect is 72 percent of what it would be without the suspicious respondents. Subtracting the attenuation ratio from 1 yields the attenuation effect in percentage terms: in this case, 28 percent.[21] At the very least, researchers planning to conduct experiments on MTurk should consider this when conducting *a priori* power analyses.

It is worth noting the similar pattern of treatment effects (or lack thereof) among respondents using suspicious IP addresses (panel 4 in Figure 2) and respondents who are flagged as potential trolls (panel 5 in the same figure). This may be due to a similar data-generating process: respondents flagged as likely to be potential trolls may have been classified as such because of inattention, which would add noise to the data in a similar manner as random responses from someone who does not speak English or pay attention to U.S. politics. Theoretically, these bad actors pose

[21] See SI 4 for additional evidence of attenuation of effects from low-quality responses, derived from a more complicated experiment we included in Study 1. In this experiment, the presence of bad actors attenuated experimental treatment effects by an average of 10.1 percent, even with a 95 percent HIT approval rate required for qualification.

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

an even larger problem if they attempt to respond to treatments in humorous or intentionally idiosyncratic ways. This does not appear to be the case here, but we provide one potential example of such behavior in SI 4.

Finally, the last panel speaks to the potential efficacy of restricting HITs to experienced MTurk Workers, consistent with Amazon's suggested best practices (Amazon Mechanical Turk, 2019*b*). This subsample reacted the most strongly to the treatment, with a 15-point average difference in evaluations of unemployment between experimental conditions (95 percent CI: [10.1, 19.8]) and a 10-point difference in evaluations of inflation (95 percent CI: [4.9, 15.2]). We are unable to fully parse different explanations for this group's responsiveness, though. On the one hand, people who have been in the MTurk pool long enough to complete more than 1000 HITs—most of which are not surveys—may be especially detail-oriented, and they may react more strongly to treatments because they read more closely. On the other hand, if they take lots of surveys, they may have developed a sense for research hypotheses and may try to respond in a hypothesis-consistent fashion. When considering qualifications based on HITs completed, experimenters using MTurk may thus face a tradeoff between respondents' attentiveness and the pitfalls of "professional survey responding." But the rate of bad actors among Workers with relatively low HIT counts—upward of 30 percent, according to Table 5—tips the scales in favor of using such a qualification.

6 Discussion and conclusion

Our studies demonstrate significant data quality problems on MTurk. In 2018, 25 percent of the data we collected were potentially untrustworthy, and that number held roughly constant into July 2020 (27 percent). The apparent contours of the problem shifted over that time period, though: the presence of duplicated and foreign IP addresses fell by about 50 percent, but the rate of apparently insincere responding—or "trolling"—rose by 200 percent. This is worth noting because the effects of these particular low-quality responses are less clear. While random responding due to satisficing, poor understanding of English, or ignorance of American politics—the latter two coming from respondents from other countries masquerading as U.S. adults—is liable to add significant noise to data collected, insincere and/or humorous responding may follow systematic patterns and thus add bias. Our studies also suggest that "problematic" respondents on the platform respond differently to experimental treatments than other subjects. Specifically, we find that bad behavior (responses originating from suspicious IP addresses or those which indicate potential trolling) adds noise to the data, which attenuates treatment effects—in the case of our experiment, by 28 percent.

Current data quality may be poor, but what's the prognosis? Since concerns about a "bot panic" surfaced in the summer of 2018, Amazon 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*a*, 2019*b*). While these measures may catch some of the worst offenders, we believe that platform's strategic incentives will cause Worker quality to continue to decline as participants devise new ways to game the system. The end result is that MTurk may indeed be emerging as a "market for lemons" in which bad actors eventually crowd out the good.

As it stands, researchers committed to paying a fair wage on MTurk may do just as well contracting with an established survey sampling firm. While we found some evidence of suspicious responding in samples curated by SSI (now Dynata) and Lucid, these rates are a fraction of their apparent analogs on MTurk. More importantly, according to our cost-benefit analysis, MTurk may not even be that much more cost-effective, if at all. Studies 2 and 3 cost 7.4 and 10.1 cents per valid respondent per minute, and we conservatively estimate that researchers can

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

conduct research on Lucid (e.g., Coppock and McClellan, 2019) for just over 10 cents per minute (e.g., Busby, 2020; Graham, 2021, 2020). With the benefits of working with an established panel—the external validity from a more representative sample, being able to conduct within-subjects experiments in a single wave thanks to panel variables—we suspect that MTurk may be best used for testing item wording, piloting treatments, and other quick research tasks.

This is partly because the chances of improvement seem low unless we can craft and implement better methods to assess and incentivize quality responding on MTurk. 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. For example, run each IP through Know Your IP to identify blacklisted IPs and multiple submissions from the same IP, or make use of similar packages built for off-the-shelf use in Stata and R (see Kennedy *et al.* 2020).
- Include questions to detecting trolling and satisficing; develop new items to this end over time, so that it is harder for bad actors to work around them.
- *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 quality 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. Along the same lines, though, researchers should also stay apprised of the going rate for responses on other platforms.
- 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.
- Restrict survey participation to Workers who have a proven track record on MTurk—those who have completed a certain threshold of HITs. Suspicious and insincere responding fell off noticeably among respondents with 1000 or more completed HITs in Studies 2 and 3, and others have noted 5,000 as an effective threshold (Amazon Mechanical Turk, 2019*b*). But note the tradeoff here—unless the Workers in question primarily complete non-survey tasks, researchers face a sample made entirely of highly experienced survey-takers, which raises questions about professional responding, demand effects, and the degree to which

https://doi.org/10.1017/psrm.2021.57 Published online by Cambridge University Press

findings generalize to people who do not see public opinion instruments on a regular basis. (This, e.g., might be especially troubling in studies involving political knowledge.)

- Importantly, research contexts that preclude the collection of IP addresses present unique challenges for scholars using MTurk. For example, researchers in the European Union are barred from collecting such information by the General Data Protection Regulation (GDPR). Not only do these researchers face significant unknown unknowns regarding their MTurk samples, but such rules may even incentivize bad actors to seek out MTurk surveys fielded in these contexts. Ultimately, the strength of professionally managed online survey panels is their selective recruiting of panelists based on known demographic characteristics. In contexts that preclude researchers from collecting IP addresses, these firms may do especially well (in terms of the relative quality of their product and in a prospective economic sense). But note that researchers who cannot collect IP addresses are not completely fumbling in the dark. It is possible to devise items that can parse respondents in one's desired sampling frame from those masquerading as someone else (e.g., asking purported Americans to write the date in MM/DD/YYYY format, showing purported Brits a picture of an elevator [or "lift"] or trash can ["bin"] and asking them to name it). Similarly, the low-incidence screener battery may be adopted across contexts to identify potentially insincere responses.
- Note that we have only touched on a handful of particular data quality issues—respondents posing as someone they are not or offering insincere responses. Other issues, including attentiveness (Thomas and Clifford, 2017), exist—and researchers should constantly be mindful of emerging threats to survey data quality.

Ultimately, researchers ought to recognize that MTurk is liable to be more prone to "lemon" responses for two reasons. First and foremost, unlike other online survey panels, MTurk does not recruit respondents based on known characteristics, thereby increasing the likelihood of obtaining respondents masquerading as someone else. Not only do MTurk Requesters know nothing about these respondents in advance, but they make up thousands of independent employers rather than one central management system. This means that the only signal of response quality 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.

Recognizing these specific issues may just be the tip of the iceberg. 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. But we believe that we are currently reckoning with a new problem in our relationship with research participants—one that demands we add "respect for validity of data" to the framework that guides this relationship. We do not believe this call is inconsistent with respect for persons, beneficence, or justice. In particular, a solution starts with researchers thoughtfully gathering data: using more credible alternatives to MTurk when possible, especially making use of pre-existing variables in online panels to leverage within-subjects studies' statistical power; using respondent requirements thoughtfully on MTurk; and, ultimately, being clear about the expectations of respondents when obtaining their consent. With these broad principles, we believe that researchers can recruit good-faith participants while fairly avoiding—and screening out, when necessary—those who contribute to the data quality problem.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2021.57.

Acknowledgements. 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 Ethan Busby, Don Green, Stephen Goggin, Matt Graham, Jonathan Nagler, John Sides, Andrew Thompson, and the anonymous reviewers for the useful comments and suggestions. Previous versions of this paper were 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.

References

Ahler DJ and Broockman D (2018) The delegate paradox: why polarized politicians can represent citizens best. *Journal of Politics* **80**, 1117–1133.

Ahler DJ and Goggin SN (2019) How does one recognize #FakeNews? Assessing competing explanations using a conjoint experiment. In *Annual Meeting of the Midwest Political Science Association*. Chicago: Midwest Political Science Association.

Akerlof GA (1970) The market for "lemons": quality yncertainty and the market mechanism. *Quarterly Journal of Economics* **84**, 488–500.

Amazon Mechanical Turk (2019*a*) MTurk worker quality and identity. Available at https://blog.mturk.com/mturk-worker-identity-and-task-quality-d3be46d83d0d.

Amazon Mechanical Turk (2019*b*) Qualifications and worker task quality. Available at https://blog.mturk.com/qualifications-and-worker-task-quality-best-practices-886f1f4e03fc.

Aronow PM, Kalla J, Orr L and Ternovski J (2020) Evidence of rising rates of inattentiveness on Lucid in 2020. Preliminary memo: https://osf.io/preprints/socarxiv/8sbe4/.

Bai H (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.

Bartels LM (2002) Beyond the running tally: partisan bias in political perceptions. *Political Behavior* **24**, 117–150.

Berinsky AJ, Huber GA and Lenz GS (2012) Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* **20**, 351–368.

Bisgaard M (2015) Bias will find a way: economic perceptions, attributions of blame, and partisan motivated reasoning during crisis. *The Journal of Politics* **77**, 849–860.

Busby EC (2020) Perceptions of extremism in the American public and elected officials. Unpublished manuscript.

Campbell DT and Stanley JC (1963) *Experimental and Quasi-Experimental Designs for Research*. Boston: Hought Mifflin Company.

Casler K, Bickel L and Hackett E (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**, 2156–2160.

Chandler J, Sisso I and Shapiro D (2020) Participant carelessness and fraud: consequences for clinical research and potential solutions. *Journal of Abnormal Psychology* **129**, 49–55.

Coppock A and McClellan OA (2019) Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics* **6**, 1–14.

Cor MK and Sood G (2016) Guessing and forgetting: a latent class model for measuring learning. *Political Analysis* **24**, 226–242.

Cornell D, Klein J, Konold T and Huang F (2012) Effects of validity screening items on adolescent survey data. *Psychological Assessment* **24**, 21–35.

Dreyfuss E (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 M, Sood G, Stone DF and Wallace J (2018) What drives demand for media slant? Unpublished manuscript. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract˙id=3009791.

Goodman JK, Cryer CE and Cheema A (2012) Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* **26**, 213–224.

Graham M (2020) When good citizens are good partisans: attributing responsibility for the COVID-19 pandemic. Unpublished manuscript.

Graham M (2021) "We Don't Know" Means "They're Not Sure." Forthcoming at *Public Opinion Quarterly.*

Hauser DJ and Schwarz N (2016) Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* **48**, 400–407.

Hillygus DS, Jackson N and Young M (2014) Professional respondents in non-probability online panels. In Callegaro M, Baker R, Bethlehem J, Göritz AS, Krosnick JA and Lavrakas PJ (eds), *Online Panel Research: A Data Quality Perspective*, New York: John Wiley & Sons, pp. 219–237.

Horton JJ, Rand DG and Zeckhauser RJ (2011) The online laboratory: conducting experiments in a real labor maret. *Experimental Economics* **14**, 399–425.

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

Kennedy R, Clifford S, Burleigh T, Waggoner P and Jewell R (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-everything-not-just-venezuela/?utm˙term=.8945c0926825.

Kennedy R, Clifford S, Burleigh T, Waggoner PD, Jewell R and Winter N (2020) The shape of and solutions to the MTurk quality crisis. *Political Science Research & Methods* **8**, 614–629.

Krosnick J (1991) Response strategies for coping with the cognitive demands of attitude meaures in surveys. *Applied Cognitive Psychology* **5**, 213–236.

Krosnick JA, Narayan S and Smith WR (1996) Satisficing in surveys: initial evidence. *New Directions for Evaluation* **70**, 29–44.

Laohaprapanon S and Sood G (2018) Know Your IP. Available at https://github.com/themains/know_your_ip.

Litman L (2019) Best recruitment practices: working with issues of non-naivete on MTurk. Available at https://www.cloudresearch.com/resources/blog/best-recruitment-practices-working-with-issues-of-non-naivete-on-mturk/.

Lopez J and Hillygus DS (2018) Why so serious? Survey trolls and misinformation. In *Annual Meeting of the Midwest Political Science Association*. Chicago. Unpublished manuscript.

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

Mitchell RE (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**, 112–119.

Mullinix KJ, Leeper TJ, Druckman JN and Freese J (2015) The generalizability of survey experiments. *Journal of Experimental Political Science* **2**, 109–138.

Mummolo J and Peterson E (2019) Demand effects in survey experiments: an empirical assessment. *American Political Science Review* **113**, 517–529.

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

Paolacci G, Chandler J and Ipeirotis PG (2010) Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* **5**, 411–419.

Paolacco G and Chandler J (2014) Inside the Turk: understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science* **23**, 184–188.

Peer E, Vosgerau J and Acquisti A (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* **46**, 1023–1031.

Pontin J (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 JP (2014) Inaccurate estimation of disparities due to mischevious responders: several Suggestions to assess conclusions. *Educational Researcher* **43**, 171–185.

Roush CE and Sood G (2020) A gap in our understanding? Reconsidering the evidence for partisan knowledge gaps. Unpublished manuscript. Available at https://www.gsood.com/research/papers/partisan˙gap.pdf.

Ryan TJ (2018) Data contamination on MTurk. Available at http://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/.

Savin-Williams RC and Joyner K (2014) The dubious assessment of gay, lesbian, and bisexual adolescents of add health. *Archives of Sexual Behavior* **43**, 413–422.

Sears DO (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**, 515–530.

Shet V (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 KA and Clifford S (2015) The generalizability of survey experiments. *Computers in Human Behavior* **77**, 184–197.

Thomas KA and Clifford S (2017) Validity and Mechanical Turk: an assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* **77**, 184–197.

Thompson AI and Busby EC (2020) Different (race) cards in the deck: directness and denials in racial messaging. Unpublished manuscript.

Vannette DL and Krosnick JA (2014) A comparison of survey satisficing and mindlessness. In Ie A, Ngnoumen CT and Langer EJ (eds.), *The Wiley Blackwell Handbook of Mindfulness*. Malden: Wiley, pp. 312–327.

Woon J (2017) Political Lie detection. Unpublished manuscript. Available at https://rubenson.org/wp-content/uploads/2017/11/woon.pdf.

Zhang C, Antoun C, Yan HY and Conrad FG (2020) Professional respondents in opt-in online panels: what do we really know? *Social Science Computer Review* **38**, 703–719.

Cite this article: Ahler DJ, Roush CE, Sood G (2021). The micro-task market for lemons: data quality on Amazon's Mechanical Turk. *Political Science Research and Methods* 1–20. https://doi.org/10.1017/psrm.2021.57

Supporting Information

SI 1 Detailed Information About MTurk Samples

SI 1.1 Descriptive Statistics on Suspicious IP Addresses

pollname	1	2	3	4	5	6	9
August 2018 Study	1885	20	13	4	1	1	1
June 2020 Study	1424	33	3	0	0	1	0
July 2020 Study	396	5	1	0	0	0	0

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

pollname	Canada	India	Other	United States	Venezuela
August 2018 Study	6	17	54	1870	42
June 2020 Study	0	12	5	1488	0
July 2020 Study	0	0	3	406	0

Table SI 1.2: Country of Origin

pollname	Buffalo	Chicago	Kansas City	Las Vegas	Los Angeles	Maracaibo	New York
August 2018 Study	77	0	28	0	44	31	72
June 2020 Study	0	40	0	35	52	0	31
July 2020 Study	0	0	0	0	0	0	0

Table SI 1.3: City of Origin

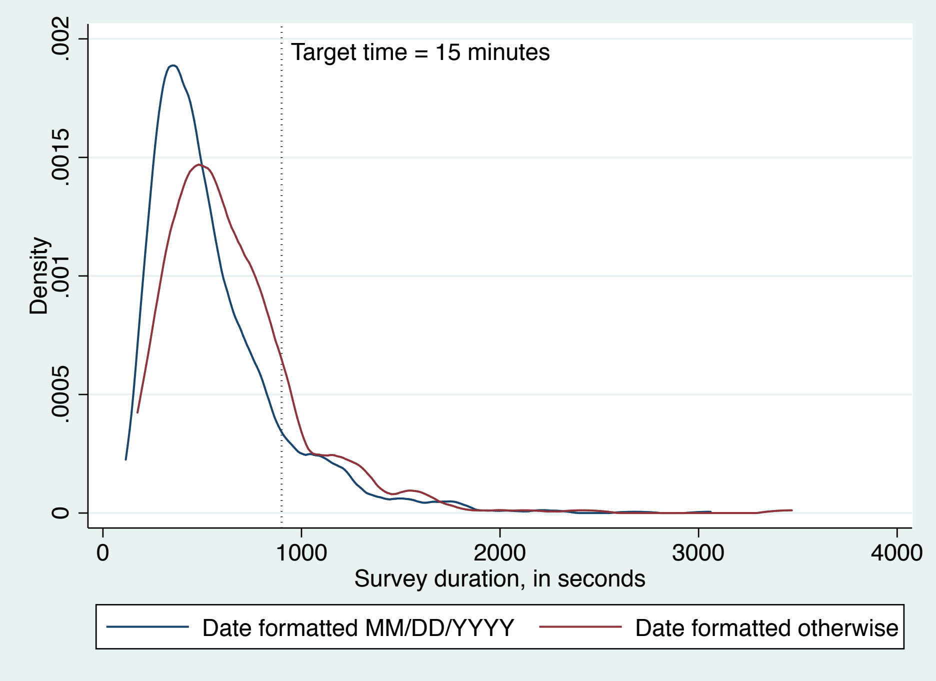
SI 1.2 Probing Duplicate IP Addresses

Duplicate IP addresses present two potential, unique problems. First, and most obviously, the same person may take the survey multiple times. Second, people may take the survey from the same network (e.g., a college campus or a workplace), which especially presents problems if these people are alerting each other to the survey—at a minimum, such a data-generating process will yield larger standard errors than those we calculate naïve to clustering. We make use of our June 2020 survey to assess the degree to which these various processes contribute to the presence of duplicated IP addresses in our data. We index each duplicated IP address and look at the start and end times of each survey from that address to do so. In particular, if the start time of survey j from address i is within ten minutues of survey $j-1$, we classify the duplicates as likely coming from the same individual. (Nearly all of these cases have end/start times within 1-2 minutes of each other.) If there are overlapping start and end times between the duplicated responses, we classify the duplicates as likely coming from a coordinated cluster of individuals. (Note, though, that this could also reflect one individual taking the survey multiple times concurrently on multiple devices.) Accordingly, we estimate that of the 37 duplicate IP addresses in the data, 10 (27%) reflect people filing mulitiple submissions and an additional 21 (57%) coming from coordinated clusters. And even with the remaining 16% of responses from duplicated IP addresses, there is likely significant heterogeneity within those clusters that should be accounted for in analysis.

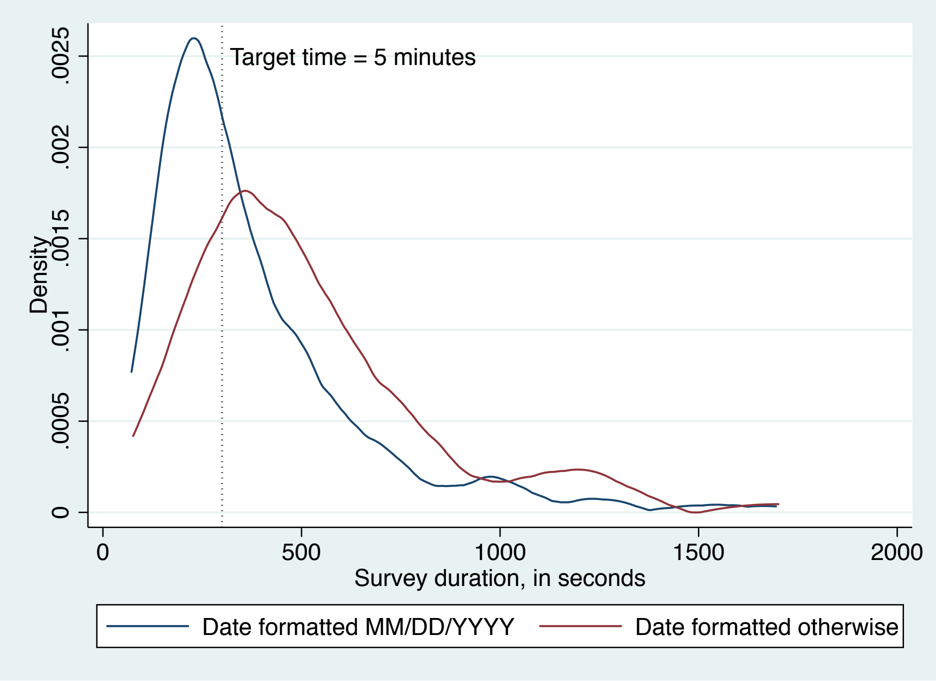
SI 1.3 Probing Foreign IP Addresses

One possibility worth investigating is whether foreign respondents actually try to take the survey genuinely—or, at the very least, spend time reading it. One might expect that people with limited English and/or understanding of American politics would speed through the survey, since they are likely taking it purely for the reward. But this is not what we find. When we use writing the date in DD/MM/YYYY format (asked on the 2020 surveys) as a proxy for being outside the U.S., we find that respondents who write the date in the non-U.S. format actually take longer, on average. The figures below show this, plotting the distributions of completion times by whether the date was written DD/MM/YYYY or MM/DD/YYYY. We find that people who use the U.S. date format complete the survey more quickly—Kolmogorov-Smirnov tests conclude the probability that these sets of response times were drawn from the same distribution is less than .001 for each survey. This suggests that respondents outside the U.S. may actually be trying to read and respond to American MTurk surveys in some meaningful way, despite the fact that they are outside the sampling frame (and are thus undesirable as survey respondents for researchers of U.S. politics). There could be other explanations—slower internet connections, for example—but one possibility is that these respondents take surveys as genuinely as possible so as to avoid detection.

Figure SI 1.1: Surveys from Likely Foreign Respondents Take Longer



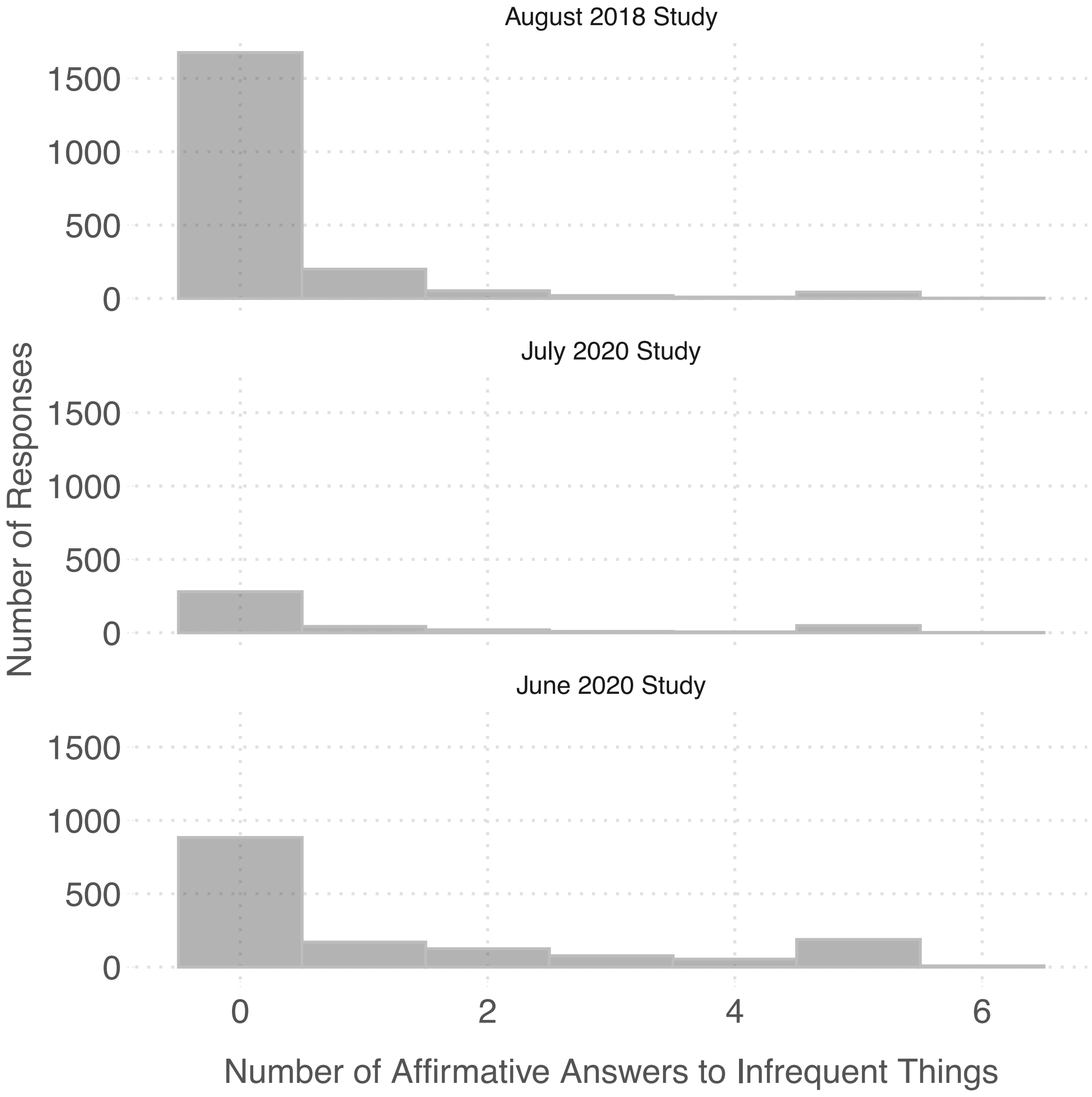
(a) June 2020



(b) July 2020

SI 1.4 Distributions of Counts of Affirmative Responses to Low-Incidence Screener Items

Figure SI 1.2: *Distribution of Counts of Affirmative Responses to Low-Incidence Screener Items Across Surveys*



SI 1.5 Additional Information Regarding Speedy Respondents

Following the same procedure for classifying slow and fast outliers described in the paper, we estimated the proportions of each in Studies 2 and 3.

In Study 2, roughly 6% of the sample is classified as fast outliers, having completed the survey in 232 seconds or less. Roughly 8% of the sample is classified as slow outliers, taking more than 1,112 seconds to finish the survey. While respondents flagged as potential trolls or originating from a suspect IP address are not any less likely to be classified as slow outliers than non-flagged respondents (7.2% vs. 7.6% of the sample, respectively, p(diff)=0.773)) they are significantly less likely to be classified as fast outliers than non-flagged respondents (4.7% vs. 7.1% of the sample, respectively, p(diff)=0.057)).

In Study 3, 14.7% of the sample is classified as fast outliers, having completed the survey in 177 seconds or less. 8.6% are classified as slow outliers—those who took longer than 799 seconds to take the survey. In this survey, 3.5% of respondents flagged as potential bad actors (by virtue of being potential trolls or for having taking the survey from a suspicious IP address) were classified as fast outliers, a figure that dwarfs in comparison to the proportion of non-suspicious respondents classified as such (roughly 18%, p(diff)=0.000)). Suspect respondents are not statistically more likely than non-suspect respondents to be classified as slow outliers (roughly 12% vs. 7.4%, diff, respectively, p(diff)=0.159).

Contrary to conventional wisdom, we do not find that respondents who are extraordinarily fast in completing the survey provide low-quality data. Table SI 1.4 models evaluations of the unemployment and inflation rates as a function of the experimental treatment (described in 5), being a fast outlier, and the interaction of the treatment with fast outlier status. The fact that the coefficients on *Out-party treatment * fast* are not substantively or statistically significant at conventional levels suggests that fast outliers do not respond differently to our experiment than respondents not classified as such. (We find similar results for our other experiment detailed in SI 4.4.) It is for this reason that we do not consider fast outliers as a source of low quality data in our broader analysis.

Table SI 1.4: *Impact of Fast Completion Times on Treatment Effects - June 2020 Survey*

	Unemployment DV	Inflation DV
Out-party treatment	-0.097***	-0.063***
	(0.016)	(0.017)
Fast	0.012	-0.004
	(0.047)	(0.049)
Out-party treatment * fast	0.011	0.043
	(0.064)	(0.067)
Constant	0.792***	0.711***
	(0.011)	(0.012)
Observations	1,425	1,425
R-squared	0.027	0.010

Standard errors in parentheses.
***p<0.01, **p<0.05, *p<0.1, two-tailed.

SI 2 Question Wording

SI 2.1 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

SI 2.2 Sincerity Self-Report

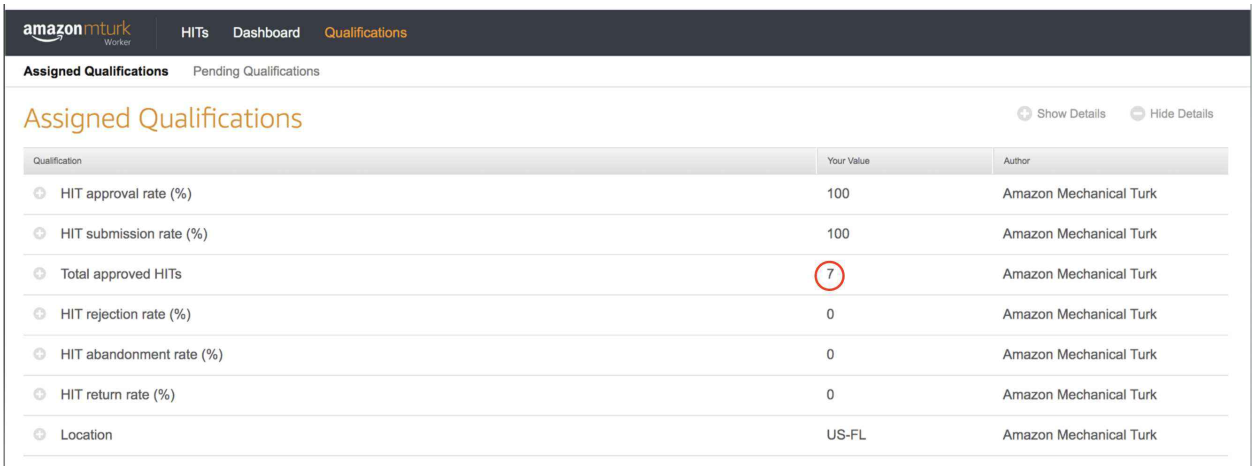
Finally, we sometimes find people don't always take surveys seriously, instead of providing humorous or insincere responses to questions. How often do you do this?

- Never
- Rarely
- Some of the time
- Most of the time
- Always

SI 2.3 Self-Reported Number of HITs Completed

We'd like to know a little more about your participation on MTurk.

To answer the question below, please visit worker.mturk.com/qualifications/assigned and look for your "Total Approved HITs" number (see graphic below). If you cannot find this information, just provide us with your best guess.



Assigned Qualifications

Qualification	Your Value	Author
HIT approval rate (%)	100	Amazon Mechanical Turk
HIT submission rate (%)	100	Amazon Mechanical Turk
Total approved HITs	7	Amazon Mechanical Turk
HIT rejection rate (%)	0	Amazon Mechanical Turk
HIT abandonment rate (%)	0	Amazon Mechanical Turk
HIT return rate (%)	0	Amazon Mechanical Turk
Location	US-FL	Amazon Mechanical Turk

About how many HITs have you completed on MTurk?

- Fewer than 100 HITs
- Between 100 and 500 HITs
- Between 500 and 1,000 HITs
- More than 1,000 HITs

SI 2.4 Experimental Item Wording (from Study 2 in June 2020)

Switching gears, we'd like to understand how you think various measures of the economy performed a few years ago, (when Barack Obama was president | when Republicans were in control of both Houses of Congress).

During 2016, (when Barack Obama was president | when Republicans controlled both Houses of Congress), unemployment decreased from 5.0% to 4.8%, a change of 0.2 percentage points. How would you interpret this change? Would you say that unemployment got better, stayed about the same, or got worse?

- Got better
- Stayed the same
- Got worse

In 2016, inflation also decreased from 2.1% to 1.9%, a change of 0.2 percentage points. How would you interpret this change? Would you say that inflation got better, stayed about the same, or got worse?

- Got better
- Stayed the same
- Got worse

SI 3 Experimental Effects by Subgroup

Table SI 3.5: Experimental Effects by Subgroup

	All non-suspicious respondents		All suspicious respondents		Suspicious IPs only		Potential trolls		High HIT (1000+) respondents	
	Unemployment DV	Inflation DV	Unemployment DV	Inflation DV	Unemployment DV	Inflation DV	Unemployment DV	Inflation DV	Unemployment DV	Inflation DV
Out-party treatment	-0.125***	-0.091***	-0.049**	-0.013	-0.029	0.023	-0.043*	-0.017	-0.150***	-0.101***
	(0.020)	(0.021)	(0.023)	(0.026)	(0.044)	(0.048)	(0.026)	(0.028)	(0.025)	(0.026)
Constant	0.775***	0.720***	0.820***	0.698***	0.795***	0.642***	0.819***	0.709***	0.782***	0.743***
	(0.015)	(0.015)	(0.017)	(0.018)	(0.032)	(0.035)	(0.018)	(0.020)	(0.018)	(0.019)
Observations	861	861	564	564	182	182	445	445	505	505
R-squared	0.043	0.022	0.008	0.000	0.002	0.001	0.007	0.001	0.068	0.029

Source: June 2020 study.

Standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.10, two-tailed.

SI 4 A Second Experiment Demonstrating Attenuation of Treatment Effects from Low-Quality Respondents

As an additional means 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 treatment effect attenuation. 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:

$$p_{ij} = p(y_i = j) = \begin{cases} p(y_i = -1) = p(y_i^* \leq \alpha_{-1}) \\ p(y_i = 0) = p(\alpha_{-1} < y_i^* \leq \alpha_0) \\ p(y_i = 1) = p(\alpha_0 < y_i^*) \end{cases} \tag{2}$$

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:

$$p(y_i = j) \sim \text{logit}^{-1}(\beta_k X_{ik} + \delta LQ_i + \gamma(LQ_i \times X_{ik}) + \varepsilon) \tag{3}$$

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 4.2. For ease of interpretation, we present marginal effects in Table SI 4.6, 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]).

Table SI 4.6: 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 ($n = 397$)		Non-serious respondents only ($n = 125$)	
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)	**14.0%**	**-9.7%**	*5.1%*	*-3.7%*	**7.5%**	**-5.3%**	*-4.6%*	*3.6%*
Gay (vs. straight)	**19.1%**	**-13.2%**	***9.3%***	***-6.8%***	**11.6%**	**-8.3%**	*-0.6%*	*0.4%*
Evangelical (vs. nothing)	**-5.8%**	**4.2%**	1.3%	-0.9%	-1.9%	1.4%	4.9%	-3.7%
Secular (vs. nothing)	**6.6%**	**-4.5%**	7.3%	-5.2%	5.6%	-3.9%	12.4%	-9.0%
Liberal (vs. nothing)	**9.4%**	**-6.4%**	2.6%	-1.9%	5.5%	-3.8%	*1.0%*	*-0.8%*
Conservative (vs. nothing)	**-8.3%**	**5.9%**	**-11.6%**	**9.0%**	**-12.4%**	**9.3%**	-16.4%	13.7%

Estimates in **bold** are significantly different from zero ($p < 0.05$).

Estimates in *italics* are significantly different from those in the non-suspicious respondents column ($p < 0.05$).

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 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 4.3 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.

SI 4.1 Question Wording for the "James" Experiment

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

SI 4.2 Results of Fully Specified Ordered Logit Model

Table SI 4.7: Impact of Low-Quality Responses on Treatment Effects - Full Ordered Logit

	All respondents	Suspicious IPs	Non-serious respondents
Low-quality response	-0.15	-0.11	-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 4.3 Calculating Attenuation Effects

From the data and the ordered logistic regression model specified above, 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 as a percentage. 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 4.4 Additional Information Regarding Speedy Respondents

Echoing the results presented in SI 1.5, we do not find that fast outliers react differently to experimental treatments than respondents who are neither extraordinarily fast 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.

Table SI 4.8: 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