The Shape of and Solutions to the MTurk Quality Crisis

February 7, 2019

Words: 7,607

Abstract: Amazon's Mechanical Turk (MTurk) is widely used for data collection, however, researchers recently noticed a decline in data quality, stemming from the use of Virtual Private Servers (VPSs) to fraudulently gain access to surveys restricted to US residents. Unfortunately, we know little about the scale and consequence of this fraud, and tools for social scientists to detect and prevent this fraud are underdeveloped. Analyzing 38 studies conducted on MTurk since 2013, we demonstrate that this problem has recently spiked, but is not new. Two new studies show that these respondents provide particularly low-quality data. We provide three solutions: software to identify fraud in existing datasets, an easy-to-use web application based on this software, and a method for blocking fraudulent respondents in Qualtrics surveys. We demonstrate the effectiveness of the screening procedure in a third study. Our results suggest that these fraudulent respondents provide unusually low-quality data, but can be easily identified and screened out.

Keywords: crowdsourcing, experiments, MTurk, survey research, online research

The advent of crowdsourcing platforms, such as Amazon's Mechanical Turk (MTurk), has been a boon for survey researchers. MTurk allows researchers to quickly collect data at a substantially lower cost than professional survey providers. The samples are not representative of any particular population, but they tend to be far more diverse than most common convenience samples and tend to replicate a variety of experimental and observational results [1–4], although studies requiring substantial trust in the experimenter [5] or using an overused experimental paradigms [6] are notable exceptions. Though met with skepticism by some, MTurk respondents tend to yield high-quality data when respondents are screened on reputation [7]. In many situations, MTurk samples have been found to provide higher quality data than student samples, community samples, and even some high-quality national samples [2,8–10]. For these reasons, the use of MTurk for survey research has grown dramatically across a variety of disciplines, including psychology [11,12], economics [13], public administration [14], and sociology [15]. One survey found that more than 1,200 studies were published in 2015 using the service [16], and another reported that more than 40% of studies published in two top psychology journals had at least one experiment that used MTurk [12]. Even studies that do not report the results of MTurk experiments often rely on the service to pilot experiments.

However, a major threat to MTurk data quality was uncovered in the summer of 2018. Several researchers reported suddenly finding high rates of poor quality responses. Many suspected these responses were generated either by bots (semi- or fully-automated code to automatically respond to surveys) or scripts (code that assists humans in responding more rapidly to certain types of questions) [17,18]. The problem, however, was quickly traced back to international respondents using Virtual Private Servers (VPS) — also sometimes referred to as Virtual Private Networks (VPN) or proxies — to mask their location and take surveys that were designed for US participants. The respondents who used VPS produced substantially lower

quality responses, including: nonsensical responses to open-ended questions, random answers to experimental manipulations, and suspicious responses to demographic questions [19–21]. While these studies gave a good indication of the source and severity of the current quality crisis, we still have little idea about the scale and duration of the problem or why it has spiked recently, nor have these studies provided solutions that can easily be incorporated into a researcher's standard workflow.

In this paper, we outline the scale of the quality crisis — its sources and its impact — and assess new methods and tools for ameliorating it. We begin by conducting an audit of our past studies on MTurk. Analyzing 38 surveys conducted over the past 5 years and encompassing 24,930 respondents, we find that VPS and non-US respondents have spiked in recent months, but that this problem likely traces back to substantially earlier, potentially placing thousands of studies at risk. Next, we detail the impacts of these VPS and non-US respondents on survey quality using two original studies (n = 2,010) that incorporate extensive quality checks. Consistent with previous studies, we find little evidence that bots are completing surveys in any notable number (and that bot detection tends to correspond to VPS use). We do, however, find that VPS users provide substantially worse quality data than other respondents, in terms of responses to explicit quality checks, answers to open-ended questions, and responsiveness to experimental treatments. Finally, we introduce new R and Stata packages that can be used retrospectively to remove fraudulent respondents from existing data, along with an online Shiny app for users of other statistical softwares. We also introduce a protocol that can be easily implemented in Qualtrics to prevent VPS users and international respondents from taking a survey. We provide evidence from a further study (n = 411) that this screening procedure is effective and causes minimal disruption.

Results

What is happening?

To better understand the quality crisis, we conducted an audit of 38 studies fielded by the authors since 2013, covering 24,930 respondents. All of these studies requested US respondents with at least a 95% approval rate on previous Human Intelligence Tasks (HITs). For all of the studies, we utilized IP Hub (https://iphub.info) to trace back the information on the IP from which the user accessed our surveys. We marked those participants who accessed the surveys through an international IP address (i.e. they took the survey from a non-US location, even though we selected only US respondents from MTurk) or used a VPS service to access the survey (i.e. their internet service provider suggested they were masking their location).

The results are stark. Not only did we discover a large number of respondents who were either using a VPS or were located outside of the US, but we also discovered that this was not a new phenomenon. Figure 1 shows the results of this audit, broken down by the month in which the study was conducted. Subfigure A shows the number of total respondents in each month. While we had more respondents in some months than others, in none of them did we have fewer than 150 unique respondents. Subfigure B shows that the largest number of fraudulent respondents comes in Summer/Fall 2018, when about 20% of respondents were coming either from a VPS or a non-US IP address, but *we notice a significant proportion of potential fraudulent respondents as far back as April 2015* (over 15% of respondents), and even some non-US IP addresses dating back to 2013.

MTurk verifies user location by requiring that those who sign up from the US provide bank account and tax information through Amazon payments to verify residence (https://www.mturk.com/worker/help). By scouring MTurk and country forums on Reddit and other sources, we found several ways workers circumvented these checks by, for example,

having an acquaintance or relative sign up for the account, signing up for an account while temporarily in the US, or by purchasing an account from a US resident.

But from where are these responses coming? It is impossible to track down a person's true location when they are using a VPS. Such services have strict privacy policies unless the VPS is being used to break a law. There are, however, a few clues we can use to make an educated guess. TurkPrime[21] used a test for speakers of English from India. They showed a picture of an eggplant and asked respondents to identify its name. A little over half of the fraudulent respondents using a VPS said the name was "brinjal," which is the Indian English term for the vegetable. From this evidence, they implied that the fraudulent users may have been from India. This, however, does not appear to be the entire explanation. In the studies used for our audit, a substantial number non-US users forgot to turn on their VPS prior to taking the surveys. This allowed us to see their true location. Subfigure C of Figure 1 shows the proportion from each country that contributed more than 4 responses in our audit. We find the largest proportion of international respondents are coming from Venezuela (about 18%), with the second largest coming from India (about 12%). Finally, Subfigure D of Figure 1 shows the substantial increase in both these groups since 2017.

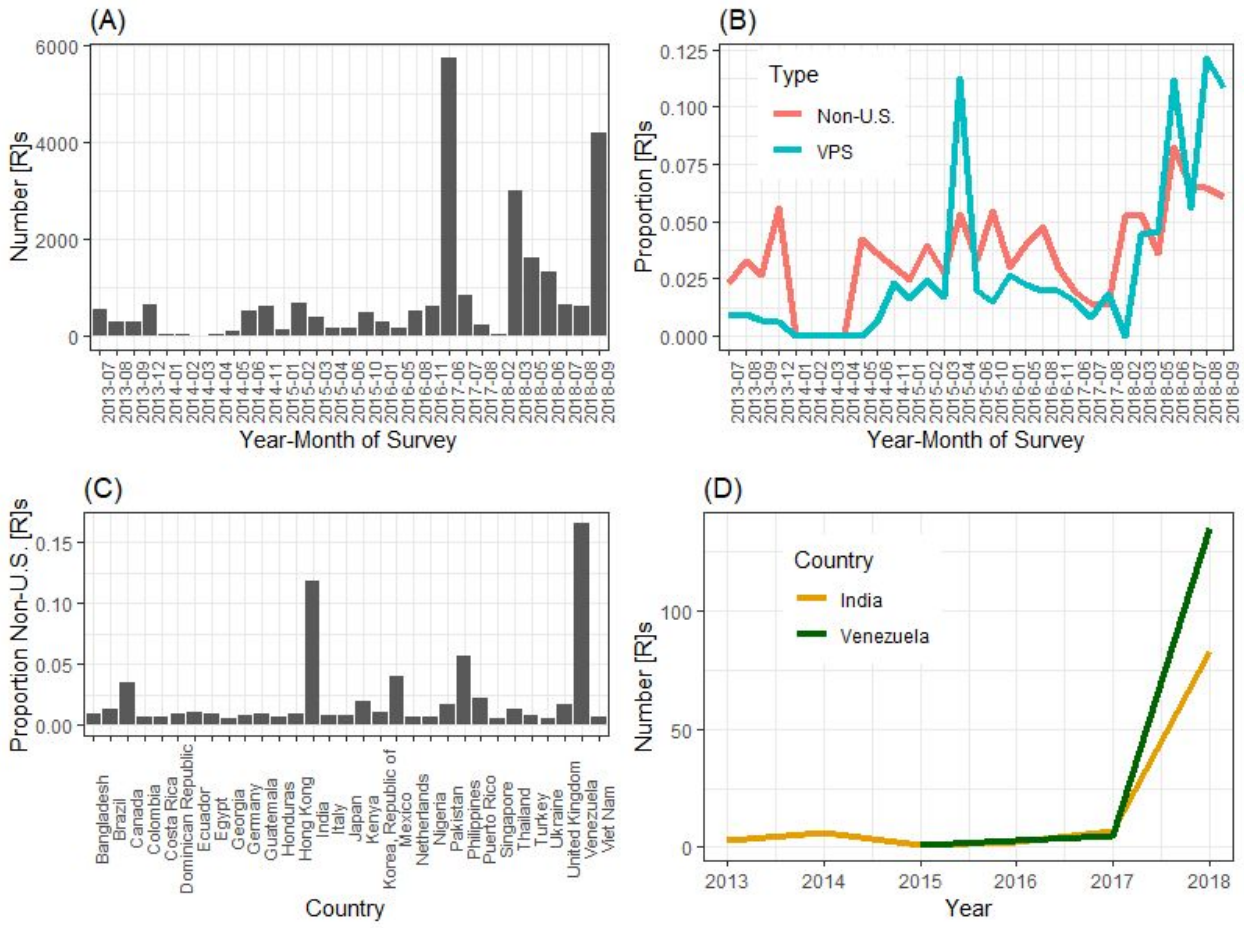

Fig. 1. Audit of past studies. These charts show the details of of 38 studies fielded by the authors since 2013, covering 24,930 respondents from MTurk. All studies requested US respondents with more than 95% approval rate on previous HITs. Subfigure A shows the number of respondents broken down by month. Subfigure B shows the proportion of respondents from a non-US IP address or using a VPS service to mask their location. Subfigure C shows the proportion of respondents who did not use a VPS to mask their location and were from a non-US location. Subfigure D shows the rise in non-US respondents from India and Venezuela since 2017.

The results in this section both raise concerns about the extent of the MTurk quality crisis and provide some indication of its likely sources. However, just because a respondent is using a VPS or is responding from outside of the US does not necessarily imply they are providing low-quality data. Many people in the US use VPSs out of privacy concerns and thus our VPS users may be valid respondents. Similarly, some US residents may be responding to

MTurk surveys while traveling or living overseas. We directly address this question in the next section.

What is the Impact?

In this section, we show the results from two studies on which we included additional quality checks to see the impact of different types of fraudulent responses on MTurk. For both studies, we used IP Hub to label IP addresses that were likely from a VPS and those that were from an international source.

Retrospective Study 1

For our first retrospective study, we recruited 576 participants from MTurk. We measured data quality in several ways, including response consistency and open-ended questions. Among the full sample, 6.8% (n = 39) provided low-quality data on at least one measure. At the end of the survey, we also utilized reCAPTCHA to weed out potential bots[22]. Six respondents completed the data quality checks on the page prior to the reCAPTCHA, but did not submit the reCAPTCHA, suggesting there may have been a very small number of bots in our survey. Five of these six respondents were using a VPS (block=1), suggesting that these potential bots can be identified using IP addresses.

Of the 576 respondents who completed the survey, 71 (12.3%) were identified as VPS users and and nine (1.6%) of uncertain status. Additionally, 38 (6.6%) were flagged for a non-US location, 25 of whom were not flagged for VPS use. Some VPS users were also located outside the US because the service they used has servers in Canada. Together, 96 (16.7%) were flagged as fraudulent, with an additional 9 (1.6%) flagged as potentially fraudulent. In the following, we refer to the remaining 81.7% who were not flagged as “valid” respondents.

We now turn to examining whether respondents whose IPs are flagged as fraudulent provide unusually low-quality data (see Figure 2). Of the valid respondents, only 2.8% (95% CI:

1.6%–4.7%) were flagged by at least one of the quality checks. Among VPS users, 23.9% (15.3%–35.5%) were flagged as providing low-quality data, while 11% (0.9%–62.6%) of respondents with an uncertain VPS status were flagged for low-quality data. Finally, among non-VPS users who were located outside of the US, 32.0% (16.0%–53.7%) were flagged as low-quality. While VPS users and foreign respondents both provided lower quality data than valid respondents ($p < .001$), data quality was indistinguishable between VPS users and foreign respondents ($p = .430$), belying the claim that many VPS users may be valid US respondents. Overall, tracing the users' IP addresses seems to be effective at identifying low-quality respondents.

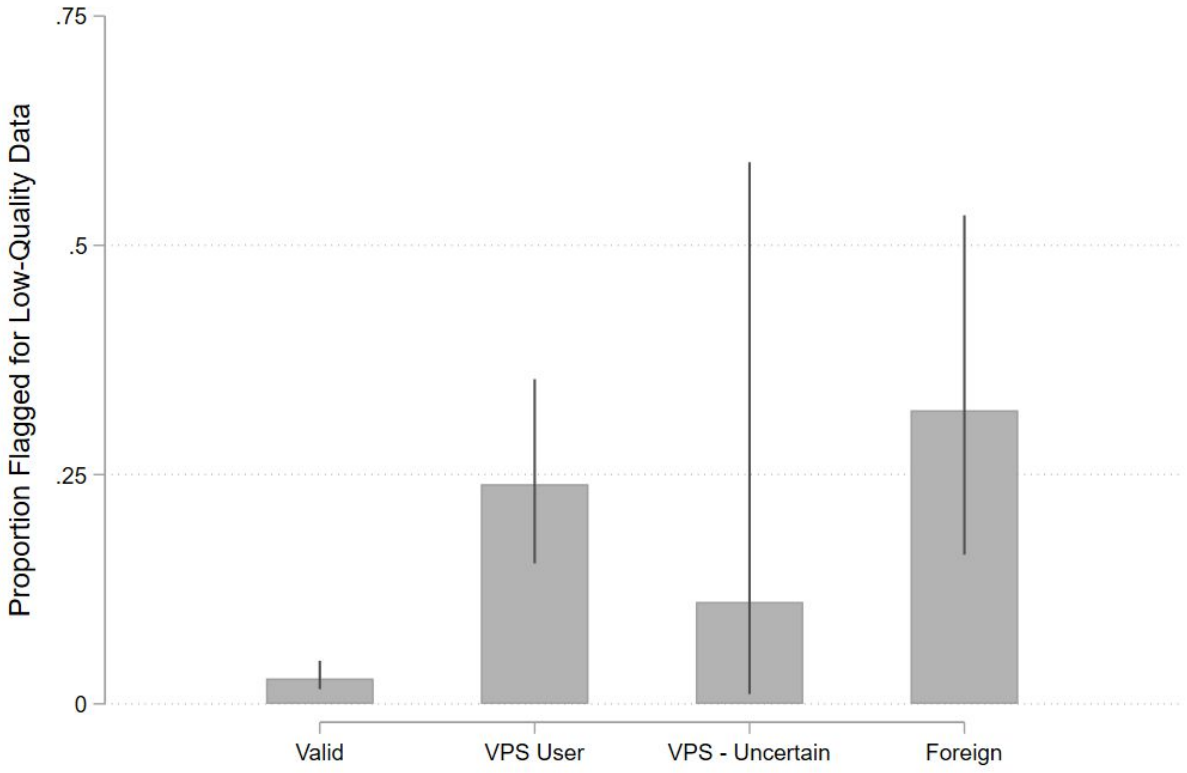

Fig. 2. Prevalence of low-quality data by respondent IP type in study 1. Proportion of participants in each VPS category failing at least one basic quality check. VPS users are those marked by IP Hub as definitely coming through a commercial server farm. Those where VPS is uncertain are those IP Hub identifies as potentially problematic, but warns of a larger likelihood of false positives. Those identified as having foreign IPs are those where the location of the connection is outside of the US.

Using additional items in the survey, we can also search for evidence of a lack of cultural knowledge that may be problematic among foreign respondents. We begin with four general political knowledge questions, which were preceded by instructions to not look up the answers[23]. Valid respondents answered 2.7 questions correctly, on average. VPS users answered significantly fewer questions correctly (2.3, p = .045). However, our other categories of fraudulent respondents did not significantly differ from the valid respondents (VPS Uncertain: M = 3.00, p = .439; Foreign: M = 2.76, p = .843). This is surprising at first glance. However, respondents can easily look up the answers to these questions and often do so[23]. Fraudulent respondents who are attempting to pass themselves off as valid may be particularly inclined to cheat. While we do not have direct measures of cheating, we can evaluate the time spent on these questions. Valid respondents spent, on average, 30 seconds on the four questions. While this may seem fast, a previous study demonstrated that likely valid responses on knowledge questions averaged about 12 seconds per question[23]. VPS users, on the other hand, spent more than four times as long (135 seconds, p < .001). Foreign respondents also spent substantially longer on the knowledge questions (81 seconds, p < .001). Only respondents with uncertain VPS status did not significantly differ from our valid respondents (47 seconds, p = .206), though this may be due to the small sample size (n = 9). This pattern of results holds even after controlling for the time spent on the remainder of the survey and a set of common covariates (education, gender, race, political interest, see Table A1), supporting the claim that this is indicative of cheating. These results suggest that our fraudulent respondents are less knowledgeable about US politics, but will put in additional effort to appear knowledgeable.

Another test involves the link between partisan identification and self-reported political ideology. The relationship should be strong among Americans, but attenuated among foreign or

inattentive respondents. Among our valid respondents, the correlation between the two variables is r = .86. However, this relationship is much weaker among VPS users (r = .45) and foreign respondents (r = .44), though not among our respondents of uncertain VPS status (r = .92). A regression analysis predicting partisan identification as a function of ideology, respondent status, and interactions between ideology and status demonstrates that ideology is a significantly stronger predictor of partisanship among valid respondents than among VPS users (p < .001) and foreign respondents (p = .003; see Table A2). Again, these results indicate that respondents who are flagged as fraudulent based on their IP addresses are less likely to have the same cultural knowledge as our valid respondents.

We also sought to more directly examine the consequences of fraudulent respondents on the substantive conclusions that would be reached by researchers. To do so, we analyze an experiment embedded in the study. Respondents in this study were asked to evaluate six target individuals based on brief vignettes and each vignette contained nine experimental conditions , plus a control condition. We estimated treatment effects among three different sets of respondents: the full sample (respondents: 576, observations: 3,456), valid respondents who are located in the US and not using a VPS (respondents: 480, observations: 2,880) and fraudulent respondents who are not located in the US or are using a VPS (respondents: 96, observations: 576). Full model details are shown in SI Table A3.

Figure 3 plots the treatment effects and confidence intervals for the valid sample on the x-axis. The left panel plots the treatment effects for the fraudulent sample on the y-axis and the right-hand panel plots the treatment effects for the full sample on the y-axis. To formalize the relationship, we regressed the nine effects estimated among the fraudulent sample on the same nine effects among the valid sample. Our null hypothesis is an intercept of 0 (no bias in treatment effects) and a slope of 1 (equal responsiveness). The constant is greater than zero (b

= .275, $p < .001$), indicating that effects are biased in a positive direction among the fraudulent subsample. The slope is much smaller than 1 ($b = .284$, $p < .001$), indicating that the fraudulent sample is less responsive to differences between the treatments (left-hand panel). We repeat this process by regressing the effects from the full sample on the effects from the valid sample. The constant is close to zero ($b = .043$, $p < .001$), indicating little bias. However, the slope is significantly smaller than 1 ($b = .871$, $p < .001$), indicating that the full sample produces smaller treatment effects (right-hand panel). These results indicate that fraudulent respondents produce substantially different treatment effects, and these respondents are prevalent enough to cause a small, but noticeable decrease in treatment effects if they are not removed from the sample. Of course, we cannot be sure of how well this finding generalizes to other studies, a question we take up in more detail in the conclusion.

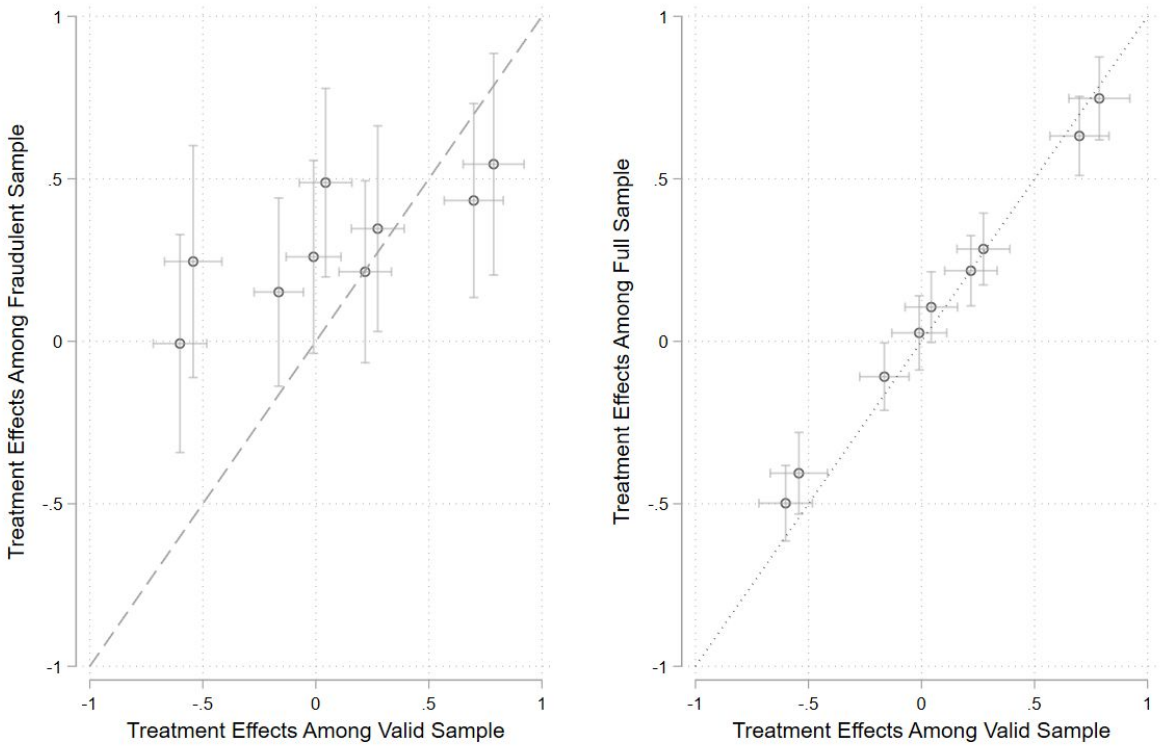

Fig. 3. Comparing treatment effects among valid and fraudulent respondents. These plots show the effect of fraudulent responses on treatment effects in the experimental part of study 1. The left-hand panel plots the estimated treatment effect among fraudulent respondents on the y-axis and the treatment effect in the valid sample on the x-axis. If the effects are the same, the dots should fall along the 45 degree dotted line. The right-hand panel plots the treatment effect including all respondents (fraudulent and valid) against the treatment effect in the valid sample.

Retrospective Study 2

In our second retrospective study, 1,641 respondents started the study and 1,440 completed it. As quality checks, we used the same five indicators from retrospective study 2. In addition, we included two more typical attention checks embedded within the experiment itself. Each followed the format of surrounding questions, but instructed participants to enter a particular response. If respondents failed the first, they were warned. If they failed the second check, they were sent to the end of the survey. These two items provide a more stringent and more common test of data quality.

We also included a reCAPTCHA at the end of the study. In this case, we had no respondents who dropped out of the survey at the reCAPTCHA page, providing no evidence for bots in our survey. Of course, it is possible that some bots failed both attention checks and were sent to the end of the survey. Nonetheless, our data is again inconsistent with bots being a significant contributor to data quality concerns.

Only 51.9% of the sample passed both instructed responses and 16.8% (n = 241) failed both. Because this latter group was removed from the survey, we cannot assess their data quality on the other five measures. Among respondents who passed both instructed responses, 13.6% were flagged as providing low-quality data according to the five alternative indicators, while 18.9% of those who failed one instructed response were flagged.

Of the 1,440 respondents who completed the survey, including those who failed the attention checks, 73.1% (n = 1,053) were valid respondents who were not flagged for using a VPS or being out of country. The remaining respondents consisted primarily of VPS users (19.3%, n = 278), followed by respondents with foreign IP addresses (6.9%, n = 100), and finally, those of uncertain VPS status (0.6%, n = 9).

Respondents whose IPs were flagged were significantly more likely to fail the attention checks, as shown in Figure 4. While 58.7% (55.7%–61.6%) of valid respondents passed both attention checks, this figure was much lower for VPS users (31.1% [25.8%–36.8%]), users with foreign IPs (41.0% [31.7%–51.0%]), and respondents of uncertain VPS status (22.2% [3.9%–67.0%]). Both VPS users and foreign respondents were significantly less likely to pass attention checks ($p < .001$), but were indistinguishable from each other ($p = .165$), again contrary to concerns that VPS users may be valid US respondents. While standard attention checks clearly help remove fraudulent responses, they are not a perfect solution. The proportion of fraudulent respondents drops from 26.9% to 20.5% when excluding respondents who failed

at least one attention checks. This figure falls only to 17.3% when removing respondents who failed either attention check. Thus, typical screeners help identify fraudulent respondents, but do not catch them all.

Turning to the five quality checks used in the previous study, 15.6% (n = 185) were flagged on at least one item, but this varies by IP type. Among valid respondents, only 8.6% [6.9%–10.5%] were flagged for low-quality data. This rate is much higher for VPS users (48.2% [40.6%–55.9%]), users with foreign IPs (31.9% [21.8%–44.0%]) and users of uncertain VPS status (25.0% [4.1%–72.4%]). While VPS users and foreign respondents both provided lower quality data than valid respondents, VPS users actually provided lower quality data than foreign respondents. Once again, our IP-based measure is effective at picking out low-quality respondents. Interestingly, we still find significant differences across these categories when restricting the sample to those who passed both attention checks, suggesting that common attention checks alone are insufficient.

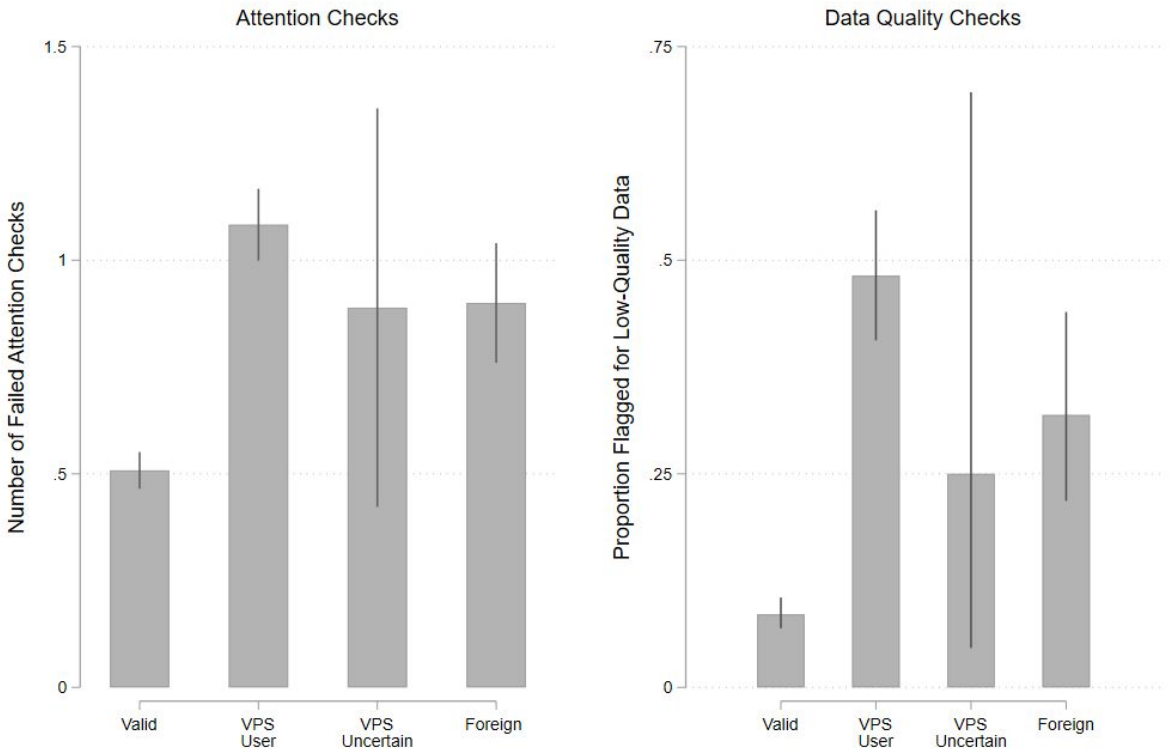

Fig. 4. Prevalence of low-quality data by respondent IP type in study 2. The left-hand panel shows the number of attention checks (instructed response) failed on average in each group. The right-hand panel shows the proportion of responses flagged for low quality in each category, who passed at least one of the attention checks.

We also examined cultural knowledge in this study by testing the relationship between partisan identification and political ideology. Again, the two variables are strongly correlated among valid respondents (r = .84). However, this relationship plummets among VPS users (r = .30) and foreign respondents (r = .45), though it remains high among the small number respondents with uncertain VPS status (r = .95). Once again, a regression model shows that ideology is more strongly associated with partisanship among valid respondents than among VPS users (p < .001) and foreign respondents (p < .001; see Table A2).

Review of Retrospective Studies

Our two retrospective studies support some of the concerns about fraudulent respondents on MTurk, while allaying other concerns. We do find clear evidence that a large number of respondents are using VPSs and that a smaller number are accessing the study from outside the US without using a VPS. However, contrary to some concerns, we found little evidence that bots make up a significant proportion of these fraudulent respondents. Consistent with the concerns of many, we found that these fraudulent respondents provide much lower quality data than respondents located in the US who are not using a VPS. These findings were consistent across a wide variety of measures, including standard attention checks, data consistency checks, open-ended responses, and measures of cultural knowledge. Notably, data quality among VPS users was consistently indistinguishable from or worse than data quality among foreign respondents, contravening the idea that many VPS users may be valid US respondents. Perhaps most importantly, fraudulent respondents were less responsive to experimental manipulations, diluting estimated treatment effects. Crucially, however, even a rate of fraud of 17% did not change the substantive conclusions of our experiment.

Detecting and Preventing Fraudulent Responses

In spite of using best practices for data collection on MTurk (e.g., HIT Approval > 95%, HITS Approved > 100[7]), our studies described above uncovered substantial rates of low-quality respondents. Fortunately, our IP-based measure was highly effective at identifying these low-quality respondents, suggesting that our measure should be incorporated into best practices. In this section, we introduce a set of tools that allow researchers to easily analyze existing datasets for fraudulent respondents and to prevent fraudulent respondents from gaining access to future surveys.

In contrast with our approach, some researchers have instead used latitude and longitude coordinates provided by survey software to identify fraudulent respondents [19,24,25]. Under this approach, responses coming from identical geographical coordinates are assumed to be stemming from a server farm used for VPS services. Supporting this method, respondents from duplicated coordinates tend to provide lower quality data. However, the mapping of an IP address to its physical coordinates is not very precise and sometimes maps IP addresses from different locations to identical coordinates [21], and respondents using less common VPS services may not be flagged for duplicate locations. For example, in the first retrospective study, 23% of VPS users had unique geographic locations and were just as likely to fail a quality check (25%) as VPS users with duplicate locations (24%). In general, the evidence linking duplicate VPS latitude and longitude to poorer quality data is weak [26]. Moreover, coordinates can only be analyzed *post hoc*, meaning they cannot be used to proactively block problematic respondents. Thus, while geographical duplicates are a potential proxy for fraudulent respondents, we recommend relying directly on IP addresses.

To assist researchers in auditing existing data, we wrote and released packages for two common programs used for statistical analysis in the social sciences -- R and Stata. The R package is available on R's Comprehensive R Archive Network (CRAN) [27] (with the most recent release on GitHub [28]). The Stata version is available from Boston College's Statistical Software Components (SSC) archive and can be installed in Stata with a single command [29]. These packages significantly streamline the process of analyzing IP addresses for researchers, from verifying IP address validity to interacting with API calls. All the researcher has to do is register for an application programming interface (API) license from IP Hub (https://iphub.info/api) to use the package. For users unfamiliar with either software, we also provide an online Shiny app that can take a comma separated values (csv) file, run the data through IP Hub, and output a csv file

to be merged with the users dataset in any statistical software [30]. These tools require minimal startup costs for most political and social scientists, as compared with other tools that require non-trivial knowledge about programming and/or API interaction [20].

While these processes offer a method for checking IP addresses after the data has been collected, it is far more efficient for both the researcher and workers if fraudulent respondents can be screened out at the beginning of the survey. We developed a method for such screening that can be easily incorporated into Qualtrics surveys. In brief, researchers need to create a free account with IP Hub (for surveys of less than 1,000 per day, or a paid account if more are expected) and add a web service call to the IP Hub API. The web service call will check the IP address against the IP Hub database and classify each respondent based on VPS use and location. The survey will then direct respondents who are using a VPS or taking the survey from abroad to an end of survey message that informs them they are ineligible for the study (see path diagram in Figure 5). Respondents whose IP status cannot be immediately identified will be provisionally allowed to take the survey, but should be checked by researchers. Just as importantly, we recommend in the protocol for researchers to warn participants that responses from outside the US are not permitted and to turn off their VPS prior to taking the survey. This warning allows respondents who may be inside the US and using a VPS for privacy reasons to turn off their VPS and continue with the survey, decreasing the number of false positives and deterring those using a VPS to commit fraud. While it is still possible for users to mask their location without use of a VPS, the methods for doing so are much more expensive in terms of time and/or money. Step-by-step instructions can be found on the Social Science Research Network (SSRN) [31].

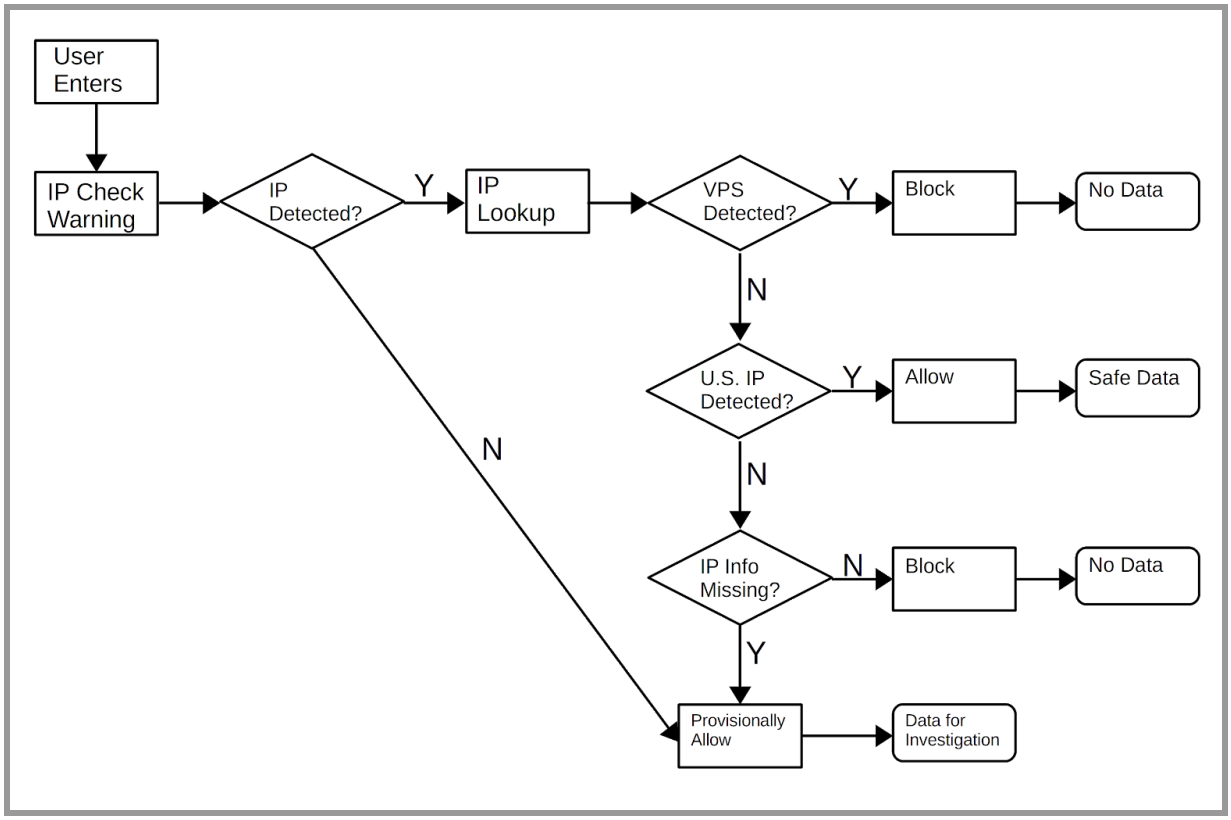

Fig. 5. Path diagram of screening protocol. This diagram shows the path diagram for our screening protocol for Qualtrics. While the protocol we provide as supplemental material is designed for a specific survey software, the diagram shown here should be adaptable to other online survey programs.

Following this protocol, we fielded a survey using Qualtrics on MTurk. We followed standard practices (US respondents, 95%+ approval rate, 100+ approved HITs) and solicited 300 HITs. We had 406 Turkers who tried to take the survey. Of those, 18 were from foreign IPs and 67 were detected as using a VPS, all of whom were successfully blocked. In six cases, we collected an IP address but were unable to collect information from IP Hub, likely because they were using very slow internet connections that did not allow the lookup process to complete. After being warned that their location would be evaluated after the study, these participants completed the survey and submitted the HIT. We checked the IP information for these participants after the data was collected, and, in all of these cases, they were found to be taking

the survey from a legitimate residential IP address in small towns in the Midwest. Because the protocol was being evaluated, we allowed an appeal process (also discussed in the full protocol) wherein they could give us their MTurk worker ID and contact us to appeal the results of the screening. We did not have any workers contact us to appeal the findings of the screening protocol. We did have one worker who complained on the survey of being a US citizen who was trying to take the survey while abroad, a claim we could not verify.

Overall, this result is quite impressive. A certain number of complaints and concerns are to be expected when working on any MTurk survey – especially if it includes attention checks. The marginal additional workload for the researcher from this protocol was minimal, while it successfully blocked access to a substantial number of respondents who would have likely contributed very low-quality data. Pre-screening respondents also has the advantage of not wasting the time of respondents who do not meet the qualifications to participate. Even for researchers who may be hesitant to add this screening to their surveys out of concern about false positives, and subsequent contact by a disgruntled worker, using the web service part of the protocol and the warning system will allow researchers to track potentially problematic responses with minimal modification of their workflow.

Discussion

While it may be tempting from some of the discussion above to conclude that MTurk is corrupted and needs to be abandoned for other platforms, this would be a mistake. MTurk is both the most popular and most studied platform for this kind of work, and shifting to other platforms would require an increase in costs that many researchers simply cannot afford. Even for scholars who can afford larger surveys, MTurk is often utilized to pilot studies prior to pre-registration and fielding. As reviewed above, MTurk samples have long provided

high-quality data that replicate many experimental and observational results, illustrating the value of the platform.

As we have seen, however, there are a few bad actors who are jeopardizing both the quality of data collected through MTurk and the future viability of the platform. Across 38 studies spanning 2013-2018, we find clear evidence that fraudulent respondents have long been on the platform, but these numbers spiked in the summer of 2018, with many of these fraudulent responses coming from India and Venezuela.

Of course, just because a respondent is using a VPS or is located outside of the US does not guarantee intentional fraud. However, across a number of tests of data quality, including common attention checks, open-ended comments, consistency checks, experimental treatment effects, and cultural knowledge, we find that these respondents tend to contribute much lower quality data and serve to diminish experimental treatment effects. Moreover, of the 85 respondents who were blocked by our screening protocol, only one contested their exclusion, suggesting that few respondents are inappropriately being flagged as fraudulent.

We provide two means to deal with this threat to data quality on MTurk. First, for studies that have already been conducted, we recommend that researchers use one of the packages we developed or the online Shiny app to identify and remove fraudulent respondents from their datasets. Because this method relies on IP addresses to identify fraud, rather than attention checks, it avoids the possibility of post-treatment bias [32]. Second, we recommend that researchers who are preparing to field a study actively check IP addresses and screen out fraudulent respondents before they have a chance to complete the study, while giving credible users, who may use a VPS for regular internet browsing, a chance to turn it off and participate. Fielding a study using this protocol, we showed that it is highly effective at screening out fraudulent respondents. Although, we should note that the protocol does not obviate the need to

use standard MTurk qualifications (95%+ HIT approval rate and 100+ approved HITs) [7]. While Amazon has been working to clean its MTurk workforce, scholars will want to maintain vigilance in the future.

Our new protocol provides a clear path forward for conducting research on MTurk, but it is less clear how to interpret studies that have already been published. Although we found evidence of fraudulent respondents as far back as 2013, rates of fraud were generally much lower prior to 2018. Moreover, a variety of replication studies conducted between 2012 and 2015 provide clear evidence of high data quality during this time frame [2,33]. Thus, it seems unlikely that fraudulent responses compromised the findings of studies prior to 2018, but it is less clear what to make of data collected more recently. Our own studies show high rates of fraudulent respondents, and these respondents contributed particularly low-quality data. However, our analysis of an experiment suggests we would reach the same substantive conclusions, though with somewhat reduced effects sizes, from these studies regardless of whether or not the fraudulent respondents were included. Of course, we have little basis for extrapolating from our experiment here to the wide variety of studies that are fielded on MTurk. Certain types of studies might be more vulnerable to the influence of fraudulent respondents, such as correlational studies assessing low or high base-rate phenomena [34] or other types of observational studies or studies using observed moderators. Bias may be particularly likely in studies of rare populations, attitudes, or behaviors, as fraudulent respondents may make up a disproportionate share of these rare categories [35,36]. For this reason, we encourage researchers to use our tools to reanalyze data they have collected on MTurk.

More generally, while this study has focused on MTurk, as the most popular crowdsourcing site for social science studies, some of the problems identified here are unlikely to be limited to this platform. The fraudulent Turkers showed a surprising level of ingenuity to

get around MTurk's standard location checks. If scholars simply moved *en masse* to a different platform, such issues are likely to simply move with them (if they have not already). This opens up a new field for scholars using crowdsourcing for their studies that combines the survey skills for the study itself with the cybersecurity understanding that is needed for online systems management. As we have seen throughout the internet, there will always be those willing to cheat the system, but even a small amount of vigilance can substantially increase the cost of this behavior, deter most potential bad actors, and minimize the damage to research.

Methods

Tracing IP Addresses

Our analyses above relied on a commercial product called IP Hub, though other alternatives are available. We find several advantages to using IP Hub. First, it is specifically targeted towards identifying likely VPS use. Other services use a much broader definition of suspicious IPs when creating their blacklist. For example, IPVOID (http://www.ipvoid.com/), used by Know Your IP and in a working paper by [20], collects its blacklist from a range of other providers and is directed towards detecting IPs potentially associated with spam, virus spread, and other behaviors. Others rely on user reports of fraud, which may not be checked by the service. Running IPVOID on the data for the second retrospective study showed that it blocked IPs from some residential providers (e.g. Comcast, AT&T, T-Mobile) and failed to block IPs from some VPS providers (e.g. DigitalOcean). Second, IP Hub's free license is relatively liberal as it allows 1,000 calls per day (30,000 per month). This compares with AbuseIPDB (https://www.abuseipdb.com/), another service used by Know Your IP, which only allows users to make 2,500 calls per month. In SI section A2, we conduct a comparison between these services and find that they generally agree with each other regarding suspicious IPs, although

there is greater correspondence between IP Hub and AbuseIPDB. We also find that IP Hub has similar performance detecting users who yielded poor quality data (true positives), while yielding fewer false positives. False positives are not necessarily an issue from a data quality standpoint, they do have the potential to burden the researcher with the task of responding to a larger number of complaints if the service is being used for screening.

IP Hub produces two levels of VPS detection. When “block” is equal to 1, it indicates that the IP was from a non-residential source, meaning high confidence that a VPS or proxy is being used. When “block” is equal to 2, it indicates that the IP is from a residential source, but has been reported as a potential VPS, meaning that there is uncertainty about whether a VPS is being used. In first section, we ignore those where the use of VPS is uncertain. In subsequent sections, we label those where “block” is equal to 2 as uncertain.

Our use of the term “fraudulent” does not imply legal liability associated with this behavior, nor that all of these respondents perceive themselves as committing fraud. Some US residents will try to take surveys while living abroad, even though they are not supposed to do so, and some US-based respondents will use VPS to mask their location out of privacy concerns. The fraudulent part stems from the attempt to claim or display a false location. As we show throughout the paper, those in both categories we label fraudulent produce, on average, much lower data quality.

Historical Studies

Our audit of previous studies includes 38 studies from the first three authors since 2013. All of these studies were approved by our home IRBs at the time of their completion: <institutions omitted for blind review>. While the studies varied in subtle ways in the qualifications requested (some requested 98%+ HIT approval, instead of 95%+), they were generally comparable in the qualifications requested and the tasks assigned were all online

social science surveys. Permission for tracing the country and VPS status of IPs in completed studies was acquired from the IRB at [author's institution] STUDY00001258.

Retrospective Studies

For our first retrospective study, we sought to recruit 575 participants from MTurk. Data was collected on August 22, 2018. While 607 respondents began the survey, only 576 respondents completed the survey and are retained for our primary analyses. Respondents were required to be located in the US, have completed at least 100 HITs, and have an approval rate greater than 95%. Respondents were paid $0.75 for completing the study. The study began with a set of demographic questions, continued to a vignette experiment involving judging the character and ideology of individuals, then on to four political knowledge questions and several questions used for quality purposes. The study was approved by the IRB of [author's institution] STUDY00000905 modification MOD00001334.

In our second retrospective study, we sought to recruit 1,400 respondents on Sept. 12-13, 2018. Though 1,641 respondents started the study, only 1,440 completed it. Respondents were required to be located in the US and have an approval rate greater than 95%. Respondents were paid $2.00 for completing the study. The study began with the verification questions, proceeded into vignettes and a conjoint experiment on trust in criminal justice algorithms, and finished with some aptitude batteries, political knowledge questions, and demographic questions. Study approval through IRB of [author's institution], study STUDY00000547 modification MOD00001380.

We measured data quality in several ways. Although researchers often use instructional manipulation checks [37,38], we do not solely rely on these because the format is easily recognizable, making it less diagnostic of attention among professional survey respondents [9]. We add novel measures of data quality that are less likely to be gamed. First, in the beginning

demographics section, we asked respondents to select the year they were born from a drop-down menu. Then in the final section of the survey we asked respondents to type in their age. Respondents whose reported age did not match their birth year were flagged as low-quality respondents. Second, we asked respondents to select their state of residence, then to report their city of residence. We expected this may be difficult for respondents from other countries and we flagged any response that was not an actual location (e.g., “Texas, Texas”) as a low-quality response. Third, at the end of the survey, we asked respondents to choose their location of residence from a multiple choice format. We piped in their responses from the beginning of the survey and the remaining ten response options were the ten least populated cities in America. We flagged any respondent who did not choose their original answer as a low-quality respondent. This check should be easy for any minimally attentive respondent, but difficult for a bot to pass. Fourth, we asked respondents to explain their main task in the survey in just a few words. Any respondent who did not provide a reasonable description of the survey (e.g., “judge people's character”) were flagged as providing low-quality data (e.g., “NICE”). Finally, we also asked respondents if they had any comments for the researcher. Although many did not answer this question, we flagged responses as low-quality if they were not in English, were unintelligible, or irrelevant to the question prompt. We then created a dichotomous variable representing whether a respondent was flagged as providing low-quality data on any of these five indicators.

In the embedded experiment for study 1, respondents each evaluated six target individuals, each of which were randomly assigned to one of 10 experimental conditions. To analyze the experiment, we stacked the data and estimated evaluations of the targets using an OLS regression model with respondent fixed effects, vignette fixed effects, dummy variables for the nine treatment conditions, and standard errors clustered on the respondent. We

re-estimated this same model among three different sets of respondents: the full sample, valid respondents, and fraudulent respondents.

Neither of these studies used IP address or data quality for decisions about compensation.

Pilot Study of Screening Technique

The pilot of the screening technique received IRB approval from [author's institution] study STUDY00001225. It was fielded on October 11, 2018. The study dealt with reactions to video and audio town halls with members of the U.S. congress. All participants were required to have a 95%+ approved HIT rating, be in the U.S., and have completed 100 HITs.

Code and Data Availability Statement

All data and code necessary for replication of study results (including all figures) will be made publicly available without restriction on Harvard's Dataverse system https://dataverse.org/ upon publication.

Works Cited

1. Clifford, S., Jewell, R. M. & Waggoner, P. D. Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics* **2**, 205316801562207 (2015).
2. Mullinix, K. J., Leeper, T. J., Druckman, J. N. & Freese, J. The Generalizability of Survey Experiments. *Journal of Experimental Political Science* **2**, 109–138 (2015).
3. Weinberg, J. D., Freese, J. & McElhattan, D. Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and Crowdsource-Recruited Sample. *Sociological Science* **1**, 292–310 (2014).
4. Berinsky, A. J., Huber, G. A. & Lenz, G. S. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Polit. Anal.* **20**, 351–368 (2012).
5. Krupnikov, Y. & Levine, A. S. Cross-Sample Comparisons and External Validity. *Journal of Experimental Political Science* **1**, 59–80 (2014).
6. Chandler, J., Paolacci, G., Peer, E., Mueller, P. & Ratliff, K. A. Using Nonnaive Participants Can Reduce Effect Sizes. *Psychol. Sci.* **26**, 1131–1139 (2015).
7. Peer, E., Vosgerau, J. & Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* **46**, 1023–1031 (2014).
8. Hauser, D. J. & Schwarz, N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav. Res. Methods* (2015). doi:10.3758/s13428-015-0578-z
9. Thomas, K. A. & Clifford, S. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Comput. Human Behav.* **77**, 184–197 (2017).
10. Anson, I. G. Taking the time? Explaining effortful participation among low-cost online survey participants. *Research & Politics* **5**, 205316801878548 (2018).
11. Paolacci, G. & Chandler, J. Inside the Turk: Understanding Mechanical Turk as a

Participant Pool. *Curr. Dir. Psychol. Sci.* **23**, 184–188 (2014).

12. Zhou, H. & Fishbach, A. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *J. Pers. Soc. Psychol.* **111**, 493–504 (2016).
13. Horton, J. J., Rand, D. G. & Zeckhauser, R. J. The online laboratory: conducting experiments in a real labor market. *Exp. Econ.* **14**, 399–425 (2011).
14. Stritch, J. M., Pedersen, M. J. & Taggart, G. The Opportunities and Limitations of Using Mechanical Turk (MTURK) in Public Administration and Management Scholarship. *International Public Management Journal* **20**, 489–511 (2017).
15. Shank, D. B. Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk. *Am. Sociol.* **47**, 47–55 (2016).
16. Bohannon, J. Psychologists grow increasingly dependent on online research subjects. *Science* | *AAAS* (2016). Available at: https://www.sciencemag.org/news/2016/06/psychologists-grow-increasingly-dependent-online-research-subjects. (Accessed: 23rd December 2018)
17. Stokel-Walker, C. Bots on Amazon's Mechanical Turk are ruining psychology studies. *New Scientist* (2018).
18. Dreyfuss, E., Barrett, B. & Newman, L. H. A Bot Panic Hits Amazon's Mechanical Turk. *Wired* (2018).
19. Dennis, S. A., Goodson, B. M. & Pearson, C. MTurk Workers' Use of Low-Cost 'Virtual Private Servers' to Circumvent Screening Methods: A Research Note. (2018). doi:10.2139/ssrn.3233954
20. Ahler, D. J., Roush, C. E. & Sood, G. The Micro-Task Market for 'Lemons': Collecting Data on Amazon's Mechanical Turk. (2018).

21. TurkPrime. After the Bot Scare: Understanding What's Been Happening with Data Collection on MTurk and How to Stop it. Available at: https://blog.turkprime.com/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it. (Accessed: 16th October 2018)
22. von Ahn, L., Maurer, B., McMillen, C., Abraham, D. & Blum, M. reCAPTCHA: human-based character recognition via Web security measures. *Science* **321**, 1465–1468 (2008).
23. Clifford, S. & Jerit, J. Cheating on Political Knowledge Questions in Online SurveysAn Assessment of the Problem and Solutions. *Public Opin. Q.* **80**, 858–887 (2016).
24. Bai, M. Evidence that A Large Amount of Low Quality Responses on MTurk Can Be Detected with Repeated GPS Coordinates. (2018). Available at: https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random. (Accessed: 5th December 2018)
25. Ryan, T. J. Data contamination on MTurk. (2018). Available at: http://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/. (Accessed: 5th December 2018)
26. TurkPrime. Understanding Geolocations and Their Connection to Data Quality. Available at: https://blog.turkprime.com/understanding-geolocations-and-their-connection-to-data-quality. (Accessed: 28th January 2019)
27. Omitted for Review. R Module: Title Omitted for Blind Review, Demonstration Available in SI. (2019).
28. Omitted for Review. *R GitHub Module: Title Omitted for Peer Review*. (Github).
29. Omitted for Review. Stata Module: Title Omitted for Blind Review, Demonstration Available in SI. *EconPapers* (2019).
30. Omitted for Review. IP Lookup Application. Available at: Demonstration available in SI.

(Accessed: 1st February 2019)

31. Omitted for Review. Full Text Available in SI: Title Omitted for Blind Peer Review. *SSRN* (2019).
32. Montgomery, J. M., Nyhan, B. & Torres, M. How conditioning on post-treatment variables can ruin your experiment and what to do about it. (2016).
33. Coppock, A. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods* 1–16 (2018).
34. Credé, M. Random Responding as a Threat to the Validity of Effect Size Estimates in Correlational Research. *Educ. Psychol. Meas.* **70**, 596–612 (2010).
35. Chandler, J. & Paolacci, G. Lie for a Dime: When most prescreening responses are honest but most study participants are impostors. *Soc. Psychol. Personal. Sci.* **8**, (2017).
36. Lopez, J. & Hillygus, D. S. Why So Serious?: Survey Trolls and Misinformation. *SSRN Electronic Journal* (2018). doi:10.2139/ssrn.3131087
37. Berinsky, A. J., Margolis, M. F. & Sances, M. W. Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *Am. J. Pol. Sci.* (2013). doi:10.1111/ajps.12081
38. Oppenheimer, D. M., Meyvis, T. & Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* **45**, 867–872 (2009).

Acknowledgements
[Omitted for blind review]