2021

Ensuring Survey Research Data Integrity in the Era of Internet Bots

Marybec Griffin
Rutgers University - New Brunswick/Piscataway

Richard J. Martino
Rutgers University - New Brunswick/Piscataway

Caleb LoSchiavo
Rutgers University - New Brunswick/Piscataway

Camilla Comer-Carruthers
Rutgers University - New Brunswick/Piscataway

Kristen D. Krause
Rutgers University - New Brunswick/Piscataway

See next page for additional authors

Quality & Quantity
https://doi.org/10.1007/s11135-021-01252-1

Ensuring survey research data integrity in the era of internet bots

Marybec Griffin[1,2] · Richard J. Martino[2] · Caleb LoSchiavo[1,2] · Camilla Comer-Carruthers[1,2] · Kristen D. Krause[1,2] · Christopher B. Stults[2,3] · Perry N. Halkitis[2,4,5,6,7]

Accepted: 28 September 2021

Abstract

We used an internet-based survey platform to conduct a cross-sectional survey regarding the impact of COVID-19 on the LGBTQ + population in the United States. While this method of data collection was quick and inexpensive, the data collected required extensive cleaning due to the infiltration of bots. Based on this experience, we provide recommendations for ensuring data integrity. Recruitment conducted between May 7 and 8, 2020 resulted in an initial sample of 1251 responses. The Qualtrics survey was disseminated via social media and professional association listservs. After noticing data discrepancies, research staff developed a rigorous data cleaning protocol. A second wave of recruitment was conducted on June 11–12, 2020 using the original recruitment methods. The five-step data cleaning protocol led to the removal of 773 (61.8%) surveys from the initial dataset, resulting in a sample of 478 participants in the first wave of data collection. The protocol led to the removal of 46 (31.9%) surveys from the second two-day wave of data collection, resulting in a sample of 98 participants in the second wave of data collection. After verifying the two-day pilot process was effective at screening for bots, the survey was reopened for a third wave of data collection resulting in a total of 709 responses, which were identified as an additional 514 (72.5%) valid participants and led to the removal of an additional 194 (27.4%) possible bots. The final analytic sample consists of 1090 participants. Although a useful and efficient research tool, especially among hard-to-reach populations, internet-based research is vulnerable to bots and mischievous responders, despite survey platforms' built-in protections. Beyond the depletion of research funds, bot infiltration threatens data integrity and may disproportionately harm research with marginalized populations. Based on our experience, we recommend the use of strategies such as qualitative questions, duplicate demographic questions, and incentive raffles to reduce likelihood of mischievous respondents. These protections can be undertaken to ensure data integrity and facilitate research on vulnerable populations.

Keywords Survey research · Internet-based research · LGBTQ research

✉ Marybec Griffin
mcg197@sph.rutgers.edu

Extended author information available on the last page of the article

Published online: 05 October 2021

1 Introduction

Internet-based research is becoming an increasingly useful tool for data collection, as it reduces the time needed to recruit participants and the number of staff needed for data collection (Das et al. 2018; McMaster et al. 2017; Schonlau and Couper 2017; Selm and Jankowski 2006). In 2020, the COVID-19 pandemic halted most in-person research activities precisely when novel research became most necessary. In order to conduct research on intersections of COVID-19 and social, economic, and health effects, numerous researchers turned to internet-based survey platforms (e.g., Qualtrics, SurveyMonkey, REDCap) to conduct research remotely. Given the future wave of peer-reviewed research using internet-based data collection methods, it is essential to ensure the integrity of these data.

Internet-based survey platforms are inexpensive, quick, and accessible ways to collect data. These mechanisms have been used to conduct research among rural populations (Campbell et al. 2018; Bowen and Ball 2020), sex workers (Thng et al. 2018; Bond et al. 2019), people who use substances (Schmidt et al. 2016; Sanchez et al. 2018), LGBTQ + populations (Guillory et al. 2018; McInroy 2016; Stults et al. 2017), and social justice movements (Harvey 2017). Internet-based research is especially useful when conducting studies among marginalized populations or those deemed "hard to reach", as they minimize burdens of travel and time (Das et al. 2018; McMaster et al. 2017; Schonlau and Couper 2017; Selm and Jankowski 2006). Furthermore, internet-based surveys offer anonymity to participants who may be less comfortable disclosing personal information or information about illegal activities (Das et al. 2018; Thng et al. 2018; Bond et al. 2019; Schmidt et al. 2016; Sanchez et al. 2018). This is especially important for adolescent populations who have not disclosed their sexual orientation or gender identity and who may live with family or friends that would not be supportive (Sterzing et al. 2018). Individuals who do not feel comfortable participating in face-to-face research are often the most underrepresented in the population of interest and internet-based research offers them an anonymous medium to participate in research as well as allowing researchers an opportunity to more fully explore the needs of the most vulnerable within marginalized populations (Sterzing et al. 2018; Russomanno et al. 2019; Iribarren et al. 2018).

Despite internal safety protocols and data protection mechanisms built into these survey platforms (i.e. ballot box stuffing, bot detection, and reCAPTCHA), it remains extremely difficult to protect against bot infiltration of online survey research (Simone 2019). Bots, defined as computer software designed to perform automated tasks for users (Nwana 1996; Eslahi et al. 2012; Teitcher et al. 2015), can be created or downloaded within minutes and deployed to complete simple automated functions or find surveys offering incentives (Yarrish et al. 2019; Godinho et al. 2020). It is important to note that this is not simply a human versus bot issue; rather human respondents are creating bots to complete surveys en masse for financial gain (Pozzar et al. 2020) as well as using other technology such as virtual private networks (VPN) and virtual private servers (VPS) to bypass safeguards against ballot box stuffing (Dennis et al. 2020). In addition to concerns about data integrity, researchers face an ethical dilemma about the accidental misuse of research funds to pay for bot responses with money reserved for participant incentives. This is an area of growing concern as federal funds for scientific research continue to decrease (Teitcher et al. 2015). Tools like reCAPTCHA offer additional protections against bot infiltrations, but are easily manipulated, and eligibility screeners can be easily deciphered with the necessary skills and time (Yarrish et al. 2019; Godinho et al. 2020).

The infiltration of bots into internet-based research is fairly commonplace and may evade detection by research staff, especially if the staff are unaware of the existence of bots, their function, and the potential impact (Yarrish et al. 2019; Godinho et al. 2020; Buchanan and Scofield 2018). Bot creation is a lucrative form of income, with incentives acting as a reward for bot creation (Yarrish et al. 2019). If data are not closely monitored, bots may complete hundreds of surveys before the activity is detected (Yarrish et al. 2019; Godinho et al. 2020; Buchanan and Scofield 2018) and may exhaust research funds allocated to incentives while leaving researchers with unusable data.

When considering both the increased reliance on internet-based surveys and the ease with which bots are created, it remains unclear whether internet-based survey platforms are helpful or harmful to research. This is especially relevant to research conducted with marginalized populations, for whom internet-based research may be more vulnerable to bot infiltration and other "mischievous responders" (Cimpian et al. 2018). Mischievous responders have been defined as survey respondents who intentionally mislead researchers by providing untruthful responses to survey items (Cimpian et al. 2018). Internet-based survey platforms may not uplift the voices of marginalized populations, instead further suppressing these voices by reducing data integrity.

Our survey was designed to understand the impact of COVID-19 on lesbian, gay, bisexual, transgender, and queer (LGBTQ+) communities in the Unites States. The purpose of this brief report is to highlight the problems associated with conducting research using internet-based survey platforms and to propose that researchers who conduct internet survey research implement a multi-faceted approach to preserving data integrity including changes to participant incentives as well as data cleaning protocols to help identify and remove bot-based responses.

2 Methods

2.1 Study design

A cross-sectional, internet-based study was conducted from May to July 2020 to understand the effects of COVID-19 on LGBTQ+ populations in the United States. Eligible participants were 18 years or older, LGBTQ-identified, and lived in the U.S or U.S territories. The survey included measures assessing demographics, substance use, sexual behavior, intimate partner violence, mental health, general health, HIV status, medication adherence, healthcare access, and COVID-19. Email addresses were recorded in a separate Qualtrics survey to facilitate delivery of incentives. The link to the incentive survey was only made available at the end of the COVID-19 survey. To ensure all COVID-19 surveys were anonymous and confidential, we were not able to collect email addresses in the same survey. The [BLINDED] Institutional Review Board approved the study protocol.

Research staff conducted internet-based recruitment in stages through professional organizations' email listservs, institutional social media accounts, and personal social media accounts. All recruitment advertising mentioned that the study was focused on the experiences of LGBTQ+ people and COVID-19 but was left intentionally vague around eligibility for the study beyond sexual orientation and gender identity to ensure that respondents could not complete the screener questions with eligible responses. Recruitment posts referenced a $5 electronic gift card for survey completion. Interested participants answered several screener questions at the beginning of the Qualtrics survey and, if

eligible, provided tacit consent prior to enrollment. After completing the COVID-19 survey, participants were directed to a separate survey to enter their email address for compensation. Participants could also complete the survey and opt out of the incentive.

2.2 Bot detection and protection

The detection of bots and the protection of data quality is a tri-fold process that includes intentional design choices for the survey, recruitment, and data cleaning steps. Separately, neither of these three steps are sufficient to reduce the number of bots; however, taken together the three processes create a strong protocol for reducing the number of bots who take internet-based surveys and removing the bot responses that bypass the other processes. For surveys, there are a number of built in survey protection settings across survey platforms. While Qualtrics survey protection settings, such as prevent ballot box stuffing (a tool that places a cookie in the browser once a person has submitted a response), reCAPTCHA (Completely Automated Public Turing Test to tell Computers and Humans Apart) scores (a question placed prior to the survey asking the respondent to identify certain items in pictures or replicate a series of letters), bot detection (a Qualtrics survey question that indicates a reCAPTCHA score that relates to the probability that the respondent is a bot), and HTTP referer verification (an option that verifies all responses come from a specific link) were activated at the launch of the survey, sophisticated bots were able to bypass these protective measures. Data were collected in two waves, before and after bot-detection. Initially, research staff conducted recruitment between May 7 and 8. Discrepancies in numbers of completed COVID-19 and incentive surveys prompted staff to pause recruitment and examine these inconsistencies, with the latter being much higher indicating that bots had specifically targeted the incentive survey. Creating two separate surveys and thereby unlinking the data from the incentive survey proved to be an effective strategy to reduce the likelihood of bots compromising the integrity of the data but offered no protection from depleting the research funds designated for participant incentives. Once the initial data were purged of bot-responses, as described below, the research team disseminated the survey again over an initial two-day trial period between June 11 and 12, using the same venues as in the initial recruitment.

2.3 Changes to recruitment

The second process to remove internet bots involves changing recruitment strategies. Prior to launching the second wave, the following protections were implemented. First, we changed the incentive structure such that, instead of a guaranteed $5 gift card, participants would enter a raffle for one of ten $100 gift cards. To prevent bots designed to identify incentive surveys, recruitment materials did not mention the gift card's value. Second, hidden questions were included throughout the survey by using the hide jQuery method in the edit question JavaScript function available through Qualtrics for the specific questions. Humans would not see the question, but bots could fill in an answer since the bots may not be designed to read JavaScript. Third, demographic questions were randomly repeated throughout the survey to check for consistency. Finally, non-U.S. IP addresses were filtered to a separate survey.

In addition to the data cleaning protocol, we developed a data integrity protocol that helped prevent bots from taking the survey as well as to identify non-human responses. The first step was to change participant incentives for completing the survey. In the first

wave of the survey, the first 200 participants were offered a $5 electronic gift card. In the second wave of the survey, participants were entered into a raffle to win one of ten $100 gift cards. Although the amount allocated for participant incentives did not change, the number of bot responses that completed the survey decreased from 633 in the first release to 23 in the second release. This change in incentive strategy was the most effective tool to disincentivize bot-based responses.

2.4 Data cleaning

The third process to protect data from internet bots was a two-part data cleaning plan.. For the COVID-19 survey, the research team developed a five-step data integrity protocol. First, we removed responses that did not complete at least 60% of the survey. Second, we used Google's invisible reCAPTCHA V3 to identify possible bots, which protects against bots and other automated programs by grading users based off a series of criteria, such as typing speed and number of requests from IP addresses (Google Developers 2020; Qualtrics 2021). We removed responses with a reCAPTCHA score of less than 0.5 from Google's reCAPTCHA V3 as suggested by the Google's developer guide (Google Developers 2020; Qualtrics 2021). Third, we removed outlier response times, defined as under 5 and over 30 min. We selected this time limits based on the average time it took study staff to completed the survey during the pilot testing of the survey. During this process, the average time for completion of the survey was 12.9 min (SD = 6.1). Fourth, we removed any responses to the qualitative survey question "How has coronavirus (COVID-19) affected your life? Please tell us as much as you feel comfortable sharing." that were exact duplicates. It should be noted that a simple logical response check is not sufficient for detecting bot responses as many bots are capable of providing logical responses to qualitative questions similar to chatbots (Augello et al. 2017). Finally, the remaining responses were checked for conflicting data. These included demographic inconsistencies (e.g., those who reported their sex assigned at birth as male and their gender identity as transgender male) and population discrepancies (e.g., those who reported living in a rural area and reported a ZIP code in a large city). Methods of detecting internal consistency, such as Even–Odd consistency and examining for straightlining (Kim et al. 2019; Ward and Pond 2015), were not used in the data cleaning process for bots because the order of answer choices for most questions were randomized.

For the incentive survey, the research staff developed a four-step process to preserve data integrity and ensure actual participants received incentives. First, we removed email address that already received compensation. Second, we removed responses that included duplicate email or IP addresses (Dennis et al. 2020). It is important to note that users were able submit multiple email addresses and use a VPN or VPS to change their IP addresses. Third, two trained research staff independently identified potential bots from the list of submitted emails, using the following protocol based on the research teams observations: the email address should not appear as random letters and the email addresses should not end in numbers exceeding four digits as these characteristics are an indication of a bot generated email address and had similar characteristics of examples from Gmail bulk account creators that can be built or bought online (Wang et al. 2017). In the final step, a third researcher checked for discrepancies between the lists and made final determinations regarding if the email was a human or bot. Gift cards were distributed to the final list of human-designated email addresses. It is important to note that eligible participants may have been incorrectly excluded from the incentive, however, participants at the beginning

of each survey were given our email address if they had any issues with the survey or receiving their expected incentive. To date, we have not received any inquiries about missing incentives.

3 Results

During the initial two-day recruitment period (Wave 1), the COVID-19 survey had a total of 1286 respondents, with 1251 (97.3%) eligible responses and 35 ineligible responses (2.7%). The incentive survey had more responses than the COVID-19 survey ($n = 1348$), an indicator that bots and/or mischievous respondents were bypassing the survey protection settings. For eligible responses, the total time to complete the survey ranged from 0.13 to 955 min ($M = 17.99$, $SD = 35.84$). Table 1 describes the changes in the number of responses as a result of the cleaning plan, across each of the steps of data cleaning for both waves of the COVID-19 survey. For the incentive survey, a total of 1348 email addresses were recorded. Table 1 describes the changes in the number of responses as a result of the cleaning plan. Overall, a total of 719 (86.0%) respondents were identified as bots and 170 (14.0%) identified as humans.

4 Discussion

This study contributes to the limited knowledge of bot behavior during internet-based survey research. As technology becomes more widely available and the average computer literacy of individuals increases, bot detection will become an increasingly larger threat to the integrity of peer-reviewed literature. Despite the threats posed by bot-based responses, researchers should not reject the idea of conducting internet based research. With additional considerations around an integrated approach to survey design, recruitment, and data cleaning that reduces the number of bots able to complete the survey and protocols to remove the bot-based responses from the data set, internet-based research remains an effective tool to engage hard to reach populations in research. This study offers a blueprint data cleaning protocol for future internet-based research studies.

Our findings indicate that the built-in data safety mechanisms are not a sufficient deterrent to bots from accessing surveys. Although we used the enhanced settings in Qualtrics, including reCAPTCHA technology, 773 (61.8%) of our survey respondents were identified as bots in our data integrity protocol. Bots were able to circumvent survey protection mechanisms such as prevention of ballot stuffing and use of HTTP referrers to protect the incentive survey. Due to the limitations of automated bot-detection software, the research team developed the following human checks for our data integrity protocol.

A large portion of our data-integrity protocol relied on a qualitative survey question where we asked the participants about other issues related to COVID-19 and the LGBTQ + community. Although we did not require an answer to this question, it proved essential in helping identify bot responses as we identified 88 (13.3%) bots in the form of exact duplicate responses that most likely would not happen by chance. However, we note that we cannot be certain that all duplicate responses were from bots. One example of an exact duplicate response was: “As a result of social isolation, coronavirus has reduced my working hours, reduced my income and shortage of living materials “, which was submitted for 11 different observations. The utility of qualitative items for bot detection is

Table 1 Data cleaning steps and remaining number of eligible observations for COVID-19 survey and incentive surveys

	Criteria	Total number of observations	Observations dropped		Observations remaining	
			N	Row %	N	Row %
Wave 1 COVID-19 survey (May 7–8, 2020)						
Step 1	Did not complete survey	1251	140	11.2	1111	88.8
Step 2	reCAPTCHA score less than 0.5	1111	190	17.1	921	82.9
Step 3	Response time less than 5 min or greater than 30 min	921	259	28.1	662	71.9
Step 4	Response to COVID qualitative question was an exact duplicate	662	88	13.3	574	86.7
Step 5	Discrepancies between ZIP Code and population size	574	57	9.9	517	90.1
Step 6	Discrepancies between working from home and identifying as an essential worker that cannot work from home	517	36	7.0	481	93.0
Step 7	Discrepancies between gender identity and sex assigned at birth	481	3	0.6	478	99.4
Wave 2 COVID-19 survey (June 11–12, 2020)						
Step 1	Did not complete survey	144	23	16.0	121	84.0
Step 2	reCAPTCHA score less than 0.5	121	2	1.7	119	98.3
Step 3	Response time less than 5 min or greater than 30 min	119	14	11.8	105	88.2
Step 4	Response to COVID qualitative question was an exact duplicate	105	0	0.0	105	100.0
Step 5	Discrepancies between ZIP Code and population size	105	7	6.7	98	93.3
Step 6	Discrepancies between working from home and identifying as an essential worker that cannot work fromhome	98	0	0.0	98	100.0
Step 7	Discrepancies between gender identity and sex assigned at birth	98	0	0.0	98	100.0
Incentive survey						
Step 1	Gift card already sent	1348	496	36.80	852	63.20
Step 2	Did not complete survey	852	16	1.88	836	98.12
Step 3	reCAPTCHA score less than 0.5	836	117	14.00	719	86.00
Step 4	Response time outliers	719	239	33.24	480	66.76
Step 5	Bot detection questions answered incorrectly	480	6	1.25	474	98.75
Step 6	Duplicate email or IP address	474	47	9.92	427	90.08
Step 7	Initial research staff screening	427	170	39.81	257	60.19

Table 1 (continued)

	Criteria	Total number of observations	Observations dropped		Observations remaining	
			N	Row %	N	Row %
Step 8	Final discrepancy check	257	140	54.47	117	45.53

confirmed by the extant literature. An online research study utilized an open-ended question with a three-character-minimum response and found no illogical responses indicative of bots. Only 17 of the 308 recorded responses "skipped" the question using spaces or periods, indicating that bots were likely unable to bypass the qualitative item (Yarrish et al. 2019). To help identify bot responses in the future, researchers should build in multiple qualitative questions and consider making one of them a requirement for submitting the survey.

Another useful check was to ask for the same data point in multiple formats. For our study, we relied on the reported ZIP Code and population size of the place where the respondent lived. Overall, we identified 57 (9.9%) bot responses from this step in our data integrity protocol. While there is the possibility for misclassification of population size by individual respondents, this will likely result in the exclusion of only a few responses. Similar to our assessment of demographic discrepancies, other researchers have used duplicate gender identity questions to trigger a logic check for subsequent data. Researchers at the University of Minnesota identified bot-based responses from individuals who indicated that they were cisgender and answered questions about trans identity, as the bots were following underlying code rather than survey logic (Perkel 2020). We recommend the use of redundant demographic measures (e.g., assessing both ZIP code and self-reported population of area) and checking for responses to "hidden" questions.

In the second wave of our data collection process, we implemented the same data integrity protocol and have only detected a minimal number of possible bots. Additionally, the change to our incentives was crucial. While a $5 gift card may not appear to be coercive, the cumulative value accrued by bots becomes a lucrative way to earn an income. Given that an average bot can be coded and repurposed to take different Qualtrics surveys or prebuilt (e.g., using websites like http://ultimatesurveybot.com/), the return on investment for a technologically savvy person is considerable. The simple change in our protocol from offering a $5 gift card for every completed survey and mentioning the amount in recruitment materials to raffling ten $100 gift cards without mentioning the amount dramatically reduced the proportion of bot responses during the initial and second two-day recruitment periods. Other studies have found that careful dissemination of participation incentives may dissuade bots, specifically if recruitment was exclusively in venues where individuals are more likely to take the survey in good faith (Yarrish et al. 2019). Furthermore, we specifically did not use social media advertising to expand the reach of our survey. While this lowers the total sample, it also helps reduce the number of bot responses by limiting the sample to a known network of LGBTQ + related virtual spaces. This is especially important when research is conducted with niche populations as bots can determine inclusion criteria responses fairly quickly (Sterzing et al. 2018; Russomanno et al. 2019; Iribarren et al. 2018). The results of our study and the available literature suggest that future internet-based research should implement raffle-based incentive structures to prevent bot infiltration.

4.1 Limitations

This study is not without limitations. The main limitation is that we could not analyze the exact effect of the change in reimbursement structure, as we had a lower number of respondents in the second wave, attributable to the large number of bot responses in the first wave. Additionally, since we utilized the same recruitment methods, most of the individuals who were interested in participating likely did so in the first wave. Despite the

lower number of human participants, the non-existence of bot-based responses is a strength of our second-wave data collection methodology.

5 Conclusion

As researchers use the internet to conduct studies with increasingly limited funding and in precarious situations, it is imperative that data integrity protocols be included in the study design from the beginning. By randomizing participant incentives and proactively building in data integrity verification questions, researchers can more effectively limit the number of and more easily identify bot responses included in the data. Furthermore, future peer-reviewed literature of internet-based research should clearly detail data integrity protocols and provide data on the identified and excluded bot responses. As internet research becomes more common, researchers across disciplines have the opportunity to create a vast collection of bot detection methodologies that serve us as researchers and science as a whole.

Acknowledgements The study was funded using Rutgers University discretionary funding and was approved by the Rutgers University Institutional Review Board.

Authors' contributions Not applicable.

Funding The study was funded using Rutgers University discretionary funding.

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest Not applicable.

Ethics approval The Rutgers University Institutional Review Board approved the study protocol.

Consent to participate All participants provided tacit consent prior to enrollment.

Consent for publication Not applicable.

References

Augello, A., Gentile, M., Dignum, F.: An overview of open-source chatbots social skills. In: International conference on internet science, pp. 236–248. Springer, Cham (2017)

Bond, K.T., Yoon, I.S., Houang, S.T., Downing, M.J., Grov, C., Hirshfield, S.: Transactional sex, substance use, and sexual risk: comparing pay direction for an internet-based US sample of men who have sex with men. Sex. Res. Soc. Policy **16**(3), 255–267 (2019). https://doi.org/10.1007/s13178-018-0366-5

Bowen, A., Ball, K.: REPORT: creating and piloting a survey to determine readiness in rural populations in Ohio (2020). https://digitalcommons.otterbein.edu/stu_doc/47. Accessed 15 Jan 2021

Buchanan, E.M., Scofield, J.E.: Methods to detect low quality data and its implication for psychological research. Behav. Res. Methods **50**(6), 2586–2596 (2018). https://doi.org/10.3758/s13428-018-1035-6

Campbell, R.M., Venn, T.J., Anderson, N.M.: Cost and performance tradeoffs between mail and internet survey modes in a nonmarket valuation study. J. Environ. Manag. **210**, 316–327 (2018). https://doi.org/10.1016/j.jenvman.2018.01.034

Cimpian, J.R., Timmer, J.D., Birkett, M.A., Marro, R.L., Turner, B.C., Phillips, G.L., 2nd.: Bias from potentially mischievous responders on large-scale estimates of lesbian, gay, bisexual, or questioning (LGBQ)-heterosexual youth health disparities. Am. J. Public Health **108**(S4), S258–S265 (2018). https://doi.org/10.2105/AJPH.2018.304407

Das, M., Ester, P., Kaczmirek, L. (eds.): Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies. Routledge, London (2018)

Dennis, S.A., Goodson, B.M., Pearson, C.A.: Online worker fraud and evolving threats to the integrity of MTurk data: a discussion of virtual private servers and the limitations of IP-based screening procedures. Behav. Res. Account. **32**(1), 119–134 (2020)

Eslahi, M., Salleh, R., Anuar, N.B.: Bots and botnets: an overview of characteristics, detection and challenges. In: 2012 IEEE International Conference on Control System, Computing and Engineering, pp. 349–354 (2012). https://doi.org/10.1109/ICCSCE.2012.6487169

Godinho, A., Schell, C., Cunningham, J.A.: Out damn bot, out: recruiting real people into substance use studies on the internet. Subst. Abuse **41**(1), 3–5 (2020). https://doi.org/10.1080/08897077.2019.1691131

Google Developers: reCAPTCHA v3: interpreting the score (2020). https://developers.google.com/recaptcha/docs/v3#interpreting_the_score. Accessed 15 Jan 2021

Guillory, J., Wiant, K.F., Farrelly, M., Fiacco, L., Alam, I., Hoffman, L., Crankshaw, E., Delahanty, J., Alexander, T.N.: Recruiting hard-to-reach populations for survey research: using Facebook and Instagram advertisements and in-person intercept in LGBT bars and nightclubs to recruit LGBT young adults. J. Med. Internet Res. **20**(6), e197 (2018). https://doi.org/10.2196/jmir.9461

Harvey, E.: The impact of Black Lives Matter on Black college students [Master's thesis, George Mason University]. Mason Archival Repository Service (2017). https://doi.org/10.13021/G8XH52

Iribarren, S.J., Ghazzawi, A., Sheinfil, A.Z., Frasca, T., Brown, W., Lopez-Rios, J., Rael, C.T., Balán, I.C., Crespo, R., Dolezal, C., Carballo-Diéguez, A.: Mixed-method evaluation of social media-based tools and traditional strategies to recruit high-risk and hard-to-reach populations into an HIV prevention intervention study. AIDS Behav. **22**(1), 347–357 (2018)

Kim, Y., Dykema, J., Stevenson, J., Black, P., Moberg, D.P.: Straightlining: overview of measurement, comparison of indicators, and effects in mail-web mixed-mode surveys. Soc. Sci. Comput. Rev. **37**(2), 214–233 (2019). https://doi.org/10.1177/0894439317752406

McInroy, L.B.: Pitfalls, potentials, and ethics of online survey research: LGBTQ and other marginalized and hard-to-access youths. Soc. Work Res. **40**(2), 83–94 (2016). https://doi.org/10.1093/swr/svw005

McMaster, H.S., LeardMann, C.A., Speigle, S., Dillman, D.A., Millennium Cohort Family Study Team: An experimental comparison of web-push vs. paper-only survey procedures for conducting an in-depth health survey of military spouses. BMC Med. Res. Methodol. **17**(1), 73 (2017). https://doi.org/10.1186/s12874-017-0337-1

Nwana, H.S.: Software agents: an overview. Knowl. Eng. Rev. **11**(3), 205–244 (1996). https://doi.org/10.1017/S026988890000789X

Perkel, J.M.: Mischief-making bots attacked my scientific survey. Nature **579**(7799), 461 (2020). https://doi.org/10.1038/d41586-020-00768-0

Pozzar, R., Hammer, M.J., Underhill-Blazey, M., Wright, A.A., Tulsky, J.A., Hong, F., Gundersen, D.A., Berry, D.L.: Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J. Med. Internet Res. **22**(10), e23021 (2020)

Qualtrics: Fraud detection—qualtrics support (2021). https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/. Accessed 15 Jan 2021

Russomanno, J., Patterson, J.G., Tree, J.M.J.: Social media recruitment of marginalized, hard-to-reach populations: development of recruitment and monitoring guidelines. JMIR Public Health Surveill. **5**(4), e14886 (2019)

Sanchez, T.H., Zlotorzynska, M., Sineath, R.C., Kahle, E., Tregear, S., Sullivan, P.S.: National trends in sexual behavior, substance use and HIV testing among United States men who have sex with men recruited online, 2013 through 2017. AIDS Behav. **22**(8), 2413–2425 (2018). https://doi.org/10.1007/s10461-018-2168-4

Schmidt, A.J., Bourne, A., Weatherburn, P., Reid, D., Marcus, U., Hickson, F., Network, T.E.: Illicit drug use among gay and bisexual men in 44 cities: findings from the European MSM Internet Survey (EMIS). Int. J. Drug Policy **38**, 4–12 (2016). https://doi.org/10.1016/j.drugpo.2016.09.007

Schonlau, M., Couper, M.P.: Options for conducting web surveys. Stat. Sci. **32**(2), 279–292 (2017). https://doi.org/10.1214/16-STS597

Simone, M.: Bots started sabotaging my online research. I fought back. STAT News (2019). https://www.statnews.com/2019/11/21/bots-started-sabotaging-my-online-research-i-fought-back/. Accessed 15 Jan 2021

Sterzing, P.R., Gartner, R.E., McGeough, B.L.: Conducting anonymous, incentivized, online surveys with sexual and gender minority adolescents: lessons learned from a national polyvictimization study. J. Interpers. Violence **33**(5), 740–761 (2018)

Stults, C.B., Kupprat, S.A., Krause, K.D., Kapadia, F., Halkitis, P.N.: Perceptions of safety among LGBTQ people following the 2016 Pulse nightclub shooting. Psychol. Sex. Orientat. Gend. Divers. **4**(3), 251–256 (2017). https://doi.org/10.1037/sgd0000240

Teitcher, J.E., Bockting, W.O., Bauermeister, J.A., Hoefer, C.J., Miner, M.H., Klitzman, R.L.: Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. J. Law Med. Ethics **43**(1), 116–133 (2015). https://doi.org/10.1111/jlme.12200

Thng, C., Blackledge, E., McIver, R., Smith, L.W., McNulty, A.: Private sex workers' engagement with sexual health services: an online survey. Sex. Health **15**(1), 93–95 (2018). https://doi.org/10.1071/SH16243

Van Selm, M., Jankowski, N.W.: Conducting online surveys. Qual. Quant. **40**(3), 435–456 (2006). https://doi.org/10.1007/s11135-005-8081-8

Wang, Z., Qin, M., Chen, M., Jia, C.: Hiding fast flux botnet in plain email sight. In: International Conference on Security and Privacy in Communication Systems, (October), pp. 182–197. Springer, Cham (2017)

Ward, M.K., Pond, S.B.: Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. Comput. Hum. Behav. **48**, 554–568 (2015). https://doi.org/10.1016/j.chb.2015.01.070

Yarrish, C., Groshon, L., Mitchell, D.M., Appelbaum, A., Klock, S., Winternitz, T., Friedman-Wheeler, D.G.: Finding the signal in the noise: miinimizing responses from bots and inattentive humans in online research. Behav. Ther. **42**(7), 235–242 (2019)

Authors and Affiliations

Marybec Griffin[1,2] · Richard J. Martino[2] · Caleb LoSchiavo[1,2] · Camilla Comer-Carruthers[1,2] · Kristen D. Krause[1,2] · Christopher B. Stults[2,3] · Perry N. Halkitis[2,4,5,6,7]

1 Department of Health Behavior, Society and Policy, Rutgers School of Public Health, Rutgers University, 683 Hoes Lane West, Piscataway, NJ 08854, USA

2 Center for Health, Identity, Behavior and Prevention Studies, Rutgers University, Piscataway, NJ, USA

3 Psychology Department, Baruch College, City University of New York, New York, NY, USA

4 Department of Biostatistics and Social and Behavioral Health Sciences, Rutgers School of Public Health, Rutgers University, Piscataway, NJ, USA

5 Rutgers Robert Wood Johnson Medical School, Rutgers University, Piscataway, NJ, USA

6 Graduate School of Applied and Professional Psychology, Rutgers University, Piscataway, NJ, USA

7 School of Public Affairs and Administration, Rutgers University, Piscataway, NJ, USA