International Journal of Social Research Methodology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tsrm20

Attack the bot: Mode effects and the challenges of conducting a mixed-mode household survey during the Covid-19 pandemic

Edanur Yazici & Ying Wang

To cite this article: Edanur Yazici & Ying Wang (2024) Attack the bot: Mode effects and the challenges of conducting a mixed-mode household survey during the Covid-19 pandemic, International Journal of Social Research Methodology, 27:6, 791-796, DOI: 10.1080/13645579.2023.2241797

To link to this article: https://doi.org/10.1080/13645579.2023.2241797

Published online: 15 Aug 2023.

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
2024, VOL. 27, NO. 6, 791–796
https://doi.org/10.1080/13645579.2023.2241797

Routledge
Taylor & Francis Group

RESEARCH NOTE

Attack the bot: Mode effects and the challenges of conducting a mixed-mode household survey during the Covid-19 pandemic

Edanur Yazici [a] and Ying Wang [b]

[a]Department of Sociology, University of Warwick, Coventry, UK; [b]Bartlett School of Planning, University College London, London, UK

ABSTRACT

Constant changes to COVID-19 restrictions have required adaptability from social scientists including responding to new challenges such as infiltration by bots. This research note presents unexpected encounters of bot infiltration and recruitment during survey data collection under pandemic conditions. The note draws from a household survey on a social housing estate in London, UK conducted in 2021. The survey investigates residents' lived experiences of the estate and housing turnover. The note discusses the limitations of online data collection, focusing on infiltration by bots and exclusion of marginalised groups. It adds to the emerging literature on bots in survey methods, making recommendations for an iterative verification and sequential multi-stage data cleaning process. It finds that online-only approaches can exclude marginalised groups. The note argues that even under pandemic conditions, face-to-face data collection can have greater reach than online only approaches. It concludes that mixed-mode household surveys can a) mitigate the challenges of a changing research environment; b) reach a broader sample; and c) provide qualitative insight for future research.

ARTICLE HISTORY
Received 14 October 2022
Accepted 20 July 2023

KEYWORDS
Bot infiltration; mixed-mode survey; household survey; Covid-19; data validity; mode effects

Introduction

Covid-19 has brought unprecedented changes to how research is conducted. Since 2020, many studies have had to be postponed or radically altered. This has required adaptability and flexibility from researchers at a time of uncertainty. Even as COVID-19 restrictions have eased, researchers have had to plan for the unexpected, building in contingencies and revising their methods. This research note presents the challenges of designing, refining, and conducting an estate wide survey under rapidly changing COVID-19 restrictions.

The survey, 'Life on Hilgrove: Better Living together' investigates residents' lived experiences on the Hilgrove Estate, focusing on the micro-geographies of residential churn understood as the movement of population within, out of, and across administrative boundaries; and intersections between churn and practices of neighbourliness, welcoming and participation on the estate and beyond. An accurate figure of population changes through churn in London's neighbourhoods, especially during the pandemic, is vital when it comes to spatial planning, service provision, budget estimates, democratic participation, and community engagement. To obtain up-to-date knowledge about what changes are taking place in this neighbourhood, we asked questions focusing on residents' tenure types, housing careers, perceptions of the estate and interactions with their neighbours. The findings feed into the Open City project which tests whether the

CONTACT Ying Wang ying-wang@ucl.ac.uk Bartlett School of Planning, University College London, London WC1H 0NN, UK

utopian ideal of the open city exists in real life and explores issues of race, migration and living with diversity in London. Hilgrove is a 1950s social housing estate with 370 dwellings. The population of the estate is younger and more ethnically diverse than the borough average (Camden Council, 2020).

In the first half of the note, we present our iterative verification and multi-stage data cleaning process, to make recommendations for responding to bot attacks. In the second half of the note, we find that an online-only phase of data collection followed by face-to-face data collection can improve the representativeness of the sample and achieve higher ethical standards than single-mode approaches. This confirms best practice with respect to mixed-mode approaches to data collection while also building in responsiveness to changing COVID-19 restrictions; improving the scale of response; complementing quantitative data with qualitative reflections from face-to-face data collection; and building relationships with respondents to support future research.

The virus and survey design

Lockdown measures and travel restrictions rendered travel for fieldwork almost impossible. Consequently, many studies changed modes of data collection, leading to a shift to online methods and tools (Nind et al., 2021). Scholars document how they transitioned to online survey methods (Burton et al., 2020); modified question design for a move from face-to-face to online-only data collection (Sastry et al., 2020) and adjusted content for switching from one mode to another (Will et al., 2020). Realising the limitations of online-only surveys, some have employed multiple socially distant modes such as postal and telephone surveys to reach those who are digitally excluded (Burton et al., 2020; Hafner-Fink & Uhan, 2021).

Drawing on existing best practice (de Leeuw, 2018; van Selm & Jankowski, 2006) while ensuring COVID-19 safety, we adopted a multi-mode approach and remained highly adaptable to COVID-19 restrictions in terms of modes and timings. Following ethical approval being granted by our institution, our fieldwork began when the 'rule of six' was still in place in England, preventing social gatherings larger than six people from no more than two households. To comply with COVID-19 rules and reduce risks for the researchers and respondents, we started with an online-first approach, complemented by a trial postal approach. We promoted the survey using social media and physical posters. We sent each address on the estate a postcard inviting residents to complete the survey with two subsequent reminder postcards. Each postcard had information about the £5 voucher incentive offered to all respondents.

Attack the block: challenges of the online survey

Following piloting and recruitment, the data collection period lasted 11 weeks. We first distributed an online version of the survey using Qualtrics (stage 1), which is the approved platform for our institution and offers a high level of adaptability with users able to use JavaScript to customise questions. Respondents were able to access the survey either via a QR code – printed on postcards and posters, or via the URL link posted on social media. To ensure that respondents were genuine residents, we included two verification measures: a required question asking for a postal address and a correctly inputted flat block name. We also trialled a sample (10%) of postal surveys to test response rates before a wider roll out. The postal survey was intended to serve as a fallback if COVID-19 restrictions were not lifted as planned for the face-to-face data collection four weeks after initial launch.

Unexpectedly, there was a massive upsurge in responses in the second week of data collection, during which we received 540 responses – more than 50 times the responses we received in the first week. This made us suspicious that our use of a publicly available survey link combined with the £5 voucher incentive had led to our survey being attacked by bots – an application that can perform and repeat a particular task faster than a human (Eslahi et al., 2012). We identified several types of response that raised suspicion of a bot attack: nonsensical responses to qualitative and demographic

questions, such as 'clean and tidy' in response to 'please describe your ethnicity'; suspicious or clearly fabricated postal addresses; patterns in responses with text-based responses being the same albeit with slight variations across 10s or 100s of responses submitted within minutes of each other; clearly randomly generated email addresses, such as a string of numbers followed by a string of letters all at the same domain name.

While bot attack detection and prevention are hot topics in cyber security, there is a limited literature on responding to bot attacks in online survey research (Ballard & Young, 2019; Godinho et al., 2020). This is despite the necessary turn to online data collection during the pandemic (Bybee et al., 2021; Griffin et al., 2021). In response to the attack, identified by clearly suspicious patterns of responses, we reviewed our verification measures and designed new exclusion criteria. We decided against a) more invasive and labour-intensive manual verification techniques such as individually contacting respondents by email for verification or b) denying respondents recognition for their contribution to study by removing the incentive or replacing it with a raffle prize draw (Teitcher et al., 2015). We introduced a captcha question; added a verification question which asked respondents to correctly identify an image of their estate out of four local social housing estates; and replaced the drop-down menu for 'which block do you live in' with a text box. We complemented this with an assessment of mode of access via QR or via weblink (assuming QR code access was more likely to come from genuine residents because QR codes were on locally displayed posters and on postcards); Internet Protocol (IP) address through which we were able to ascertain respondents' rough geographical locations; and, where respondents' privacy settings allowed, respondents' latitude and longitude at the time of completion. Once we introduced these additional verification measures, the response rate slowed rapidly, indicating the effectiveness of our bot detection techniques.

On this basis, we developed a three-stage exclusion process to systematically clean data to ensure quality and validity. In stage 1, we removed all who had failed the verification question; who submitted a clearly fabricated address; did not provide a name and postal address; did not provide a postal address and did not have a UK IP address; all duplicate responses; and all incomplete responses. In stage 2, we removed all responses that met our criteria for suspicious email and suspicious text (a randomly generated string of letters or numbers, or non-sensical entries); whose given postal address did not match their chosen flat block; and those who met the criteria for suspicious text and whose address did not appear on a list of valid addresses for the estate. In stage three we removed all responses that met the criteria for a suspicious email address and were submitted within an hour of each other and where the mode of access was via URL link.

Our approach to data verification and cleaning has similarities with Bybee et al. (2021), which suggests that there is an emergent best practice for bot detection in the social sciences. Bybee et al. (2021) also adopted a multi-stage exclusion and cleaning process which included assessing duplicate responses, incomplete responses, and assessing names and email addresses. Where we differ, however, is our introduction of additional verification methods and assessment of mode of access. Whereas Bybee et al. (2021) individually followed up with respondents by email to verify whether they were genuine, we decided against this due to concerns about: research fatigue, respondent privacy, and sensitivity for potentially vulnerable respondents. Instead, we added image verification questions and took IP addresses into account. Our iterative verification process and sequential multi-stage data cleaning process ensured greater data validity. By combining different verification methods and cross-referencing responses with available data on residential addresses on the estate, we are reasonably certain that responses included in our analysis were completed by genuine residents instead of bots. Our measures may guide those wishing to conduct similar surveys and provide an efficient way to respond to bot attacks in comparison with highly technical algorithmic detection advocated by Eslahi et al. (2012) making our approach especially appropriate for social science researchers.

A closer look at the valid online responses indicated that the online approach privileged those who are digitally literate and are confident with written English, which skewed our responses towards younger, more educated, and white British respondents. To address this, and reach a broader range of respondents, we began our second mode of data collection by collecting survey responses door-to-door.

On Hilgrove: face to face data collection

The face-to-face element of data collection began following the lifting of COVID-19 restrictions and after two of the researchers received their second dose vaccines. Nevertheless, we maintained social distancing and offered to wear masks when we were invited into respondents' homes. As most flats on the estate open onto outdoor walkways, there was sufficient natural ventilation for us to not need to wear masks while door-knocking.

We ran eight door knocking sessions and completed two rounds of door knocking on the estate. To respect respondents' privacy, the first round of door knocking excluded addresses that had already completed the survey online. The second round of door knocking excluded all residents who expressed no interest in being surveyed or had already completed a survey at the door. We varied the times and days of the door knocking sessions to elicit as broad a range of respondents as possible. This included daytime and evening sessions as well as weekday and weekend sessions. Each door knocking session was conducted by at least two researchers working in pairs to ensure researcher safety. Each researcher also kept a field diary in which they made qualitative observations and methodological reflections.

Although door knocking is more labour intensive, we achieved greater coverage of residents on the estate with higher response rates. With 111 valid responses from 370 addresses, we reached a sample of 30.0% of households, among which almost 49% were approached through door-knocking. The response rate for the first round of door-knocking was 31.0% in comparison to a 13.8% response rate in the online survey and an 8.0% response rate in the postal survey trial. This suggests that in the context of a survey targeting a social housing estate online only methods yield lower response rates in comparison to face-to-face surveys. When taking the bot attack into account, we found that systematically cleaning data was almost as labour intensive as door-knocking. Although online-only approaches are often chosen for their efficiency in terms of researcher labour-time, our experience was that conducting research face-to-face yielded more useful contextual data than dealing with the bot attack which required carefully assessing the responses we received.

More importantly, the door knocking phase was more representative of estate residents, with a higher proportion of responses from women, women from racialised minorities and residents with no formal qualifications. This suggests that although online-only methods can provide ease and cost efficiency, they are less likely to be inclusive in research settings such as ours. This suggests that it is also important to take mode effects into account during analysis. In addition, door-knocking provided more opportunities to gain insights into sensitive questions, such as ethnicity, which was asked as a self-defined question to avoid methodological nationalism (Landolt et al., 2021). In the online survey, our self-describe box led to a non-response rate of 22.0%, 14.2% higher than the sample collected at the doorstep. Moreover, we found that the face-to-face mode yielded more nuanced responses such as 'Black African Somalian', 'Albanian Kosovan Muslim' or 'Polish, Eastern European with German family'. This approach captured identities excluded by standard categorisations and centred respondent experience. As we are interested in churn and relationships to place at the international scale, the face-to-face mode has preparatory advantages for the qualitative stages of the research project. The face-to-face mode allowed us to familiarise ourselves with the estate, identify potential future participants and add qualitative insight that would not have

been captured by the online survey alone such as the relationship between perceptions of churn and flat block size.

Conclusion: advantages of mixed-mode approaches

The findings of this survey indicate that mixed-mode approaches to survey data collection remain best practice. Combining online and offline methods ensure greater adaptability in responding to changes in COVID-19 restrictions, allowing researchers to shift flexibly from one mode to another. Moreover, we have added to the emerging literature on bot attacks in survey research. We have demonstrated that in response to suspicious patterns in responses, it is necessary to a) introduce robust and multiple verification measures including captcha questions and bespoke questions that only target respondents are able to answer and b) develop multi-stage and sequential cleaning and exclusion criteria including mode of access, email format and time of survey submission relative to other responses. Finally, we find mode type influences the type of respondent, with online-only approaches yielding higher response rates from digitally literate respondents. Overall, we argue that mixed-mode approaches should persist after the pandemic.

Acknowledgement

We would like to thank our participants from the Hilgrove Estate. We are also grateful to the Editor and anonymous reviewers for their helpful feedback that have helped us to improve this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research for this article was funded by the Economic and Social Research Council (grant reference ES/T009454/1).

Notes on contributors

Edanur Yazici is a Research Fellow in the department of Sociology at the University of Warwick. Her research focuses on race, migration, and city life. She is particularly interested in the British asylum system alongside the spatial dynamics of how we live together.

Ying Wang is a Research Fellow in the Bartlett School of Planning, University College London. Her research has been concerned with social integration and social governance in the UK and China. She is committed to an interdisciplinary and mixed-method approach to urban research.

ORCID

Edanur Yazici http://orcid.org/0000-0002-0407-2664
Ying Wang http://orcid.org/0000-0002-8664-6894

References

Ballard, A. M., Cardwell, T., & Young, A. M. (2019). Fraud Detection Protocol for Web-Based Research Among Men Who Have Sex with Men: Development and Descriptive Evaluation. *JMIR Public Health Surveill*, *5*(1), e12344. https://doi.org/10.2196/12344

Burton, J., Lynn, P., & Benzeval, M. (2020). Effects of the COVID-19 crisis on survey fieldwork: Experience and lessons from two major supplements to the U.S. Panel study of income dynamics. *Survey Research Methods*, *14*(2), 241–245. https://doi.org/10.18148/srm/2020.v14i2.7746

Bybee, S., Cloyes, K., Baucom, B., Supiano, K., Mooney, K., & Ellington, L. (2021). Bots and nots: Safeguarding online survey research with underrepresented and diverse populations. *Psychology and Sexuality*, *13*(4), 1–11. https://doi.org/10.1080/19419899.2021.1936617

Council, C. (2020). Hilgrove Summary. Unpublished internal council report. London Borough of Camden.

de Leeuw, E. (2018). Mixed-mode: Past, present, and future. *Survey Research Methods*, *12*(2). https://doi.org/10.18148/srm/2018.v12i2.7402

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

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

Griffin, M., Martino, R. J., LoSchiavo, C., Comer-Carruthers, C., Krause, K. D., Stults, C. B., & Halkitis, P. N. (2021). Ensuring survey research data integrity in the era of internet bots. *Quality & Quantity*, *56*(4), 2841–2852. https://doi.org/10.1007/s11135-021-01252-1

Hafner-Fink, M., & Uhan, S. (2021). Life and attitudes of Slovenians during the COVID-19 pandemic: The problem of trust. *International Journal of Sociology*, *51*(1), 76–85. https://doi.org/10.1080/00207659.2020.1837480

Landolt, P., Goldring, L., & Pritchard, P. (2021). Decentering methodological nationalism to survey precarious legal status trajectories. *International Journal of Social Research Methodology*, *25*(2), 1–13. https://doi.org/10.1080/13645579.2020.1866339

Nind, M., Coverdale, A., & Meckin, R. (2021). *Changing social research practices in the context of covid-19: Rapid evidence review*, https://eprints.ncrm.ac.uk/id/eprint/4458/1/NCRM%20Changing%20Research%20Practices_Rapid%20Evidence%20Review_FINAL%20REPORT.pdf

Sastry, N., McGonagle, K., & Fomby, P. (2020). Effects of the COVID-19 crisis on Survey fieldwork: Experience and lessons from two major supplements to the U.S. panel study of income dynamics. *Survey Research Methods*, *14*(2), 241–245. https://doi.org/10.18148/srm/2020.v14i2.7752

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

van Selm, M., & Jankowski, N. W. (2006). Conducting online surveys. *Quality and Quantity*, *40*(3), 435–456. https://doi.org/10.1007/s11135-005-8081-8

Will, G., Becker, R., & Weigand, D. (2020). Effects of the COVID-19 crisis on survey fieldwork: Experience and lessons from two major supplements to the U.S. Panel study of income dynamics. *Survey Research Methods*, *14*(2), 241–245. https://doi.org/10.18148/srm/2020.v14i2.7753