Survey Research Methods (2020)
Vol. 14, No. 2, pp. 159-163
doi:10.18148/srm/2020.v14i2.7761

© European Survey Research Association
ISSN 1864-3361
http://www.surveymethods.org

Partnering with Facebook on a university-based rapid turn-around global survey

Frauke Kreuter et al.
Joint work of 20 authors from the following institutions: University of Maryland, University of Mannheim, IAB, Facebook, Harvard University, Stanford University, Johns Hopkins Bloomberg School of Public Health, Carnegie Mellon University

This paper describes a partnership between Facebook and academic institutions to create a global COVID-19 symptom survey. The survey is available in 56 languages. A representative sample of Facebook users is invited on a daily basis to report on symptoms, social distancing behavior, mental health issues, and financial constraints. Facebook provides weights to reduce nonresponse and coverage bias. Privacy protection and disclosure avoidance mechanisms are implemented by both partners to meet global policy and industry requirements. Country and region-level statistics are published daily via dashboards, and microdata are available for researchers via data use agreements. Over 1 million responses are collected weekly.

Keywords: COVID-19; Facebook; COVID-19 symptom survey; partnership; probability sample

1 Introduction

The COVID-19 pandemic's unprecedented economic and social dislocations mean policymakers worldwide are in urgent need of high-quality data to track the spread of the virus, to evaluate whether the interventions that promote social distancing are working, and to guide decision making about when to loosen and tighten such interventions. Yet most local, state, and national statistical systems are not designed to provide rapid (e.g., daily) updates (National Academies of Sciences, Medicine, et al., 2017). As a consequence, policymakers and researchers have sought to gather critical data from existing private information systems. The INESS national statistics office of France, for example, turned to France's bank card association for access to anonymized microdata on consumer spending to improve its forecasts of the French Gross Domestic Product during the pandemic[1]. Likewise, technology companies and cell phone service providers are offering data on population movement relevant to understanding the coronavirus crisis. Prominent examples include Facebook's and Google's mobility maps, which chart movement trends by geography and across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential.[2]

Data on individuals, including their early COVID-like symptoms, contact behaviors, and mental and financial well-being, are harder (if not impossible) to obtain. While data on mobility can guide research on the potential for contacting someone with COVID-19, they contain no information about the health statuses of those who are tracked. Only surveys can make connections between the disparate pieces of information available in organic data, e.g., the linkage of symptoms and contact behaviors at the individual level. However, in most countries, there is no unified address list available from which a random sample can be drawn to field such a survey, which often means survey data are at risk of suffering from large selection biases. This paper reports on an effort to launch a rapid worldwide COVID-19 symptom and contact behavior survey that was a joint initiative between academics and Facebook to address some of these problems. In addition to outlining the survey design, we describe the nature of this global collaboration.

How to strengthen research through partnership?

To compare the effects of different policy decisions across nations, internationally standardized measurement instruments are of great appeal. It is therefore not surprising to

Please cite this paper with the full list of authors: Frauke Kreuter, Neta Barkay, Alyssa Bilinski, Adrianne Bradford, Samantha Chiu, Roee Eliat, Junchuan Fan, Tal Galili, Daniel Haimovich, Brian Kim, Sarah LaRocca, Yao Li, Katherine Morris, Stanley Presser, Joshua A. Salomon, Tal Sarig, Kathleen Stewart, Elizabeth A. Stuart, Ryan Tibshirani

Contact information: Frauke Kreuter, Joint Program in Survey Methodology 1218 Lefrak Hall College Park, MD 20742 (E-mail: fkreuter@umd.edu)

[1] www.reuters.com/article/us-health-coronavirus-france-stats/nowcasting-the-economy-how-pandemic-forced-french-forecasters-to-get-real-idUSKBN22I1XF

[2] https://dataforgood.fb.com/tools/disease-prevention-maps/; https://www.google.com/covid19/mobility/

see researchers mounting surveys on websites and inviting people world-wide to participate[3]. The downside of such an approach is lack of control over who is exposed to the website, who is participating and how often, and—maybe most importantly—who is missed.

About 2.6 billion people use Facebook[4]. The size of this user population enabled worldwide coverage for our web-based COVID19 Symptoms Survey. Researchers from the World Health Organization, the Delphi Group at Carnegie Mellon University, the Joint Program in Survey Methodology at the University of Maryland, and a group of experts in public health and disease modeling at Harvard, Stanford, Yale, and Johns Hopkins University collaborated to provide timely data for policy development relevant to COVID-19. Having academic partners added external review of the study design and measurement as well as subject matter expertise ensuring higher data quality overall. It also allowed for implementing a data pipeline aligned with the notice-and-consent principle of contextual integrity (Nissenbaum, 2009), meaning a clear and understandable connection between data recipient and data collection purpose. Besides providing the sample and a mechanism to correct for its bias, Facebook (experimentally) optimized the survey invitation to improve click-through-rates (the ratio of users who click on a link to the total number of users who view the invitation) by 40-50%.

How to protect privacy?

The added complexity of academic researchers from different institutions not only collecting, but analyzing the survey data, necessitated special attention to privacy control, including ensuring that ineligible third parties, including researchers without a signed Data Use Agreement as well as Facebook itself, cannot access individually identified respondent answers. Carnegie Mellon University's Delphi Research Center was responsible for data collection in the United States, and the Joint Program in Survey Methodology at the University of Maryland was responsible for global data collection. Respondents to the survey are recruited via a special invitation on the top of their Facebook News Feed, which is visually distinct from the actual News Feed items. After seeing the invitation in their News Feed, people who click on the invitation see an interstitial that offers information about the survey and what data will (and will not) be shared. They then click to get to an off Facebook survey hosted by the two universities using the online survey platform Qualtrics.

Explicit consent is requested at the start of the survey on Qualtrics. A version of the survey targeted to the European Union (EU) comports with the General Data Protection Regulation (GDPR). In EU countries we also obtain explicit permission for the data to be shared with researchers outside the EU that are not named in the initial consent statement. There are no substantial differences between the two questionnaires beyond the two differing consent statements. Privacy protections do not stop at the consent statement, since notification can only be comprehensive or comprehensible, but not both (Nissenbaum, 2011). In order to balance public access needs and privacy protection, a two tiered access system is implemented. The public can access aggregated statistics via a publicly-accessible application programming interface (API). Researchers may access non-public, non-aggregated survey data only if the appropriate Data User Agreements have been executed by the institutions with which they are affiliated. Only academic and non-profit researchers with specified research goals may request access to non-public, non-aggregated survey data[5].

Facebook creates a unique identifier (Candidate Identification Number, or CID) for each user that clicks on the link to the university-hosted survey. Each CID enables Facebook to construct a flag indicating whether a sampled user responded to the survey and to link the presence of a response to their account information. This allows Facebook to construct weights that adjust for the characteristics of the survey respondents relative to the state or country, thus reducing bias due to the respondents potentially differing from the target population. Importantly, the use of the CID allows these bias correction weights to be formed without the universities sharing individual responses with Facebook and without Facebook sharing individual user data with the universities[6]. Not all CIDs created correspond to a survey response since users who navigate to the survey on Qualtrics may choose not to participate. Once daily the university survey hosts send Facebook a listing of the CIDs that correspond to a survey response, for which Facebook returns relevant weights.

What questions to ask?

Our main priority was to ensure that the survey captured the most critical real-time indicators of disease severity while remaining short (<5 minutes). The large and diverse sample required careful question adaptations to ensure that the instrument remained understandable for a wide audience in a very heterogeneous set of countries. We gathered input from subject matter experts from the World Health Organization and international non-profit organizations, as well as public health experts and physicians to guide the process of balancing the need for detailed information with the reality of what information can be collected in these circumstances. The

[3] https://covid-19data.org/

[4] Q1 earning call April 29, 2020.

[5] https://dataforgood.fb.com/docs/covid-19-symptom-survey-request-for-data-access/

[6] The only information Facebook shared with Carnegie Mellon University Delphi Research Center and the Joint Program in Survey Methodology at the University of Maryland was the Candidate Identification Number, the weight, and locale information to display the survey in the respondent's preferred language.

survey is currently fielding in 53 languages, with Facebook overseeing the translation work, though—because of time constraints—without back-translation. Respondents are initially presented with the survey in the language that matches their Facebook settings, but respondents can opt to switch to a different language. Planning for eventual translation and localization informed the questionnaire writing and design process, with a goal to ask questions using straightforward language without the use of English colloquialisms or references to complex medical terms.

The questionnaire asks about current symptoms, access to testing, testing outcome, and contacts outside of their home. Other items included self-reported household financial outlook and indicators for nervousness, depression, and anxiety, adapted from the K10 scale (Kessler et al., 2003). A 5 day "look back" period was used for mental health measures, in order to examine these constructs in a rapidly changing environment. We adapted the health-related, testing, and contact questions from Ebola, AIDS/HIV, and other COVID-19 studies. We also ask for respondents' self-reported region of residence that corresponds to first level administrative divisions such as province, state, or region. In the United States, we request the respondent's ZIP code since it requires low respondent burden, achieves a high response rate, and is easily aggregated to the state level. For other countries, the first level administrative division is self-reported using a drop-down question that provides a complete list of regions for each country. This reduces potential for misreporting[7].

How to select the sample and adjust for nonresponse?

Key to ensuring our data are useful are proper and sustainable sample selection of Facebook users, a statistical adjustment for nonresponse based on a rich set of information, and, to the extent possible, adjustments to reduce the inferential gap between the population of active Facebook users and the general population. These sampling design and bias correction procedures are the main factors that distinguish this study from an opt-in or voluntary response internet survey[8]. The first aspect of this is achieved by drawing a new, random sample of Facebook app users each day. Sampling is stratified by level 1 administrative region and a simple random sample is then drawn within each region. Of course, nonresponse and patterns of engagement with the Facebook app produce differences between the resulting sample of respondents and the Facebook population. Additionally, not everyone in every country has a Facebook account so the sampling frame does not cover the desired population of inference. As a result, a weight mentioned above is created by Facebook and added to the externally collected survey results. The Facebook weighting procedure adjusts for non-response and coverage biases so that the weighted sample better matches the population to be represented in each country according to the respondent age, gender, level 1 administrative region, and other attributes which Facebook researchers have determined correlate with survey participation. In addition, other geographical variables are added to improve spatial representation. Regularization is applied to the prediction models to minimize the variance of the weights. Finally, Facebook performed a second step of post stratification over administrative 1 region, gender and age to correct for coverage bias.

What to keep in mind when analyzing the data?

Users of the daily COVID-19 data from the API[9] should note that the current day's estimates would typically be available two days later due to the weighting and aggregation process[10]. Estimates are not reported for locations with insufficient responses. In order to minimize the impact of various sources of error, we recommend analysts focus on temporal variation. Although estimates for a single point in time may be affected by many error sources (stemming from both sample selection and measurement procedures) these errors are likely to remain constant over relatively short periods of time, thereby producing unbiased estimates of change over time (Kohler, Kreuter, & Stuart, 2019).

Conclusions

Putting all the pieces together from start to finish in barely a few weeks required not only a large team working in parallel, but also some processes conducted in parallel that ideally should be done sequentially. Luckily we had patient legal and engineering teams at the various entities, dealing with seemingly endless iterations of documents and files calling for changes in the questionnaire and data collection procedures. In hindsight it would have been useful to spell out a change process from the beginning, and to think of the costs for each change when communicating with stakeholders. Standardizing globally might only be possible with such a short questionnaire; for topics to be examined in more detail, country-specific implementation may be necessary. Survey items aren't the only part of a project like this that requires translation. Translating mindsets and languages between the disciplines, academics, corporations and government entities is also required. But our survey data would have been impossible to produce without them.

[7] International Survey Preview https://umdsurvey.umd.edu/jfe/preview/SV_2mWYHEMq5ZoUBNj?Q_CHL=preview&Q_JFE=qdg

[8] For example, mobile phone symptom survey tracking apps (e.g., "How We Feel" https://howwefeel.org/, "COVID-19 Symptom Study" https://covid.joinzoe.com/us, "COVItrackerD-19 Citizen Science" https://eureka.app.link/covid19).

[9] https://covidmap.umd.edu/api.html

[10] For the U.S. data methodology see https://cmu-delphi.github.io/delphi-epidata/api/covidcast.html

Acknowledgement

Work described here was partially funded through the NSF RAPID award ID 2028683. The following Facebook employees were also heavily involved in this project and we are extremely thankful for their support: Tali Alterman Barash, Curtiss Cobb, Andres Garcia, Andi Gros, Ahmed Isa, Alex Kaess, Faisal Karim, Ofir Eretz Kedosha, Shelly Matskel, Roee Melamed, Amey Patankar, Irit Rutenberg, Tal Salmona, David Vannette. At the University of Maryland we also thank Katie McKeon and Joseph M. Smith for their fast turnarounds on important legal and ethical steps.

References

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., … Walters, E. E., et al. (2003). Screening for serious mental illness in the general population. *Archives of general psychiatry*, *60*(2), 184–189.

Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability Sampling and Causal Analysis. *Annual Review of Statistics and Its Application*, *6*(1), 149–172. doi:10.1146/annurev-statistics-030718-104951

National Academies of Sciences, E., Medicine et al. (2017). *Innovations in federal statistics: Combining data sources while protecting privacy*. National Academies Press.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32–48.

Commentary

Kreuter et al. present a highly ambitious academia-industry partnership to globally survey COVID-19 symptoms, perceptions, and behaviors. The initiative complements an ongoing survey project in the U.S. led by researchers at Carnegie Mellon University. Its scope and temporal granularity—daily random samples of Facebook users in more than 200 countries—is hard to imagine without close collaboration between leading academic institutions and a globally operating company with access to over 30% of the world's population. The survey itself is very short, mostly comprising of items on symptoms, activities relevant for potential coronavirus exposure and spread, and of psychosocial indicators. Depth is traded for breadth in coverage and rapid operational readiness.

We applaud the principal investigators for having set up such a bold and relevant project in such a short time. In view of the vast expertise gathered, one should not expect us to say anything astute the project team would not have thought of. Still we feel that there are a couple of concerns that are worth discussing.

Clearly, the representativeness of a sample of Facebook customers of the general population is a contentious issue. This specifically applies to a global survey where user figures may differ between countries and social groups. Random sampling among users and adjustments to population controls including age group, region, and gender are an important first step to establish representativeness given potential differences in COVID prevalence and severity along these lines. However, we were wondering what the "other attributes which Facebook researchers have determined correlate with survey participation" are, whether they are known to the PIs and whether they include indicators of education and (digital) literacy. The coverage issue seems to be particularly critical in countries with low literacy rates.

Second, it is well-known that self-reports in surveys are prone to measurement error. Interestingly though, we have learned that epidemiological symptom surveys often overestimate disease prevalence due to unspecific symptoms which are easily confused with seasonal flu and allergies (e.g., https://www.wired.com/story/survey-data-facebook-google-map-covid-19-carnegie-mellon/). We suspect that social desirability drives self-reports of mobility behavior in the opposite direction, in particular, if these are queried after items about symptoms in a survey – the former could prime respondents with negative consequences of non-compliant behavior.

Both in terms of representativeness and measurement, validation is key. While individual-level validation of self-reported mobility would be feasible, at least in principle, against Facebook's own behavioral data from mobile apps, this would raise data privacy issues and would contradict the project's current policy to decouple the survey from Facebook user data. Still we wondered whether there would be ways to tap paradata from Facebook profiles while at the same time preserving the privacy of sensitive health information collected from customers through the survey. At the less sensitive level of U.S. counties, researchers at Carnegie Mellon University validated self-reported symptoms from Facebook and Google surveys against Google searches for specific symptoms, medical tests, and doctor visits. In the present initiative, area-level validation is facilitated in countries with comprehensive COVID testing, while it will become more complex in other countries with poorer health systems and data.

Finally, we were thinking about other problem areas that might profit from such a partnership. Quite an obvious candidate is the spread of misinformation about COVID in social networks. Facebook has acknowledged the problem, and has taken steps for its containment (https://about.fb.com/news/2020/05/coronavirus/). An academia-industry partnership in researching and combating misinformation about COVID may lend additional insight, transparency, and credibility to this effort.

Simon Munzert
Hertie School of Governance
Berlin

Peter Selb
University of Konstanz

Reply to Munzert and Selb

These are very thoughtful comments and valuable suggestions.

The survey frame undoubtedly suffers from some of the coverage biases of all web surveys involving literacy and access to the internet. We believe the potential error in self-reports is less of a concern given our primary interest in changes over time as such error is unlikely to vary over relatively short periods of time. Nevertheless, additional research efforts are underway to better understand these issues.

With the help of an NSF RAPID-Award (ID 2028683), several items in the survey were synchronized with the probability based surveys in the Understanding America Study (https://uasdata.usc.edu/). The World Bank also included several comparable items in their world-wide telephone data collections. These and other data collections can serve as resources for methodological assessments, and the kind of validation you describe.

As of May 2020 The Social Experts Action Network (SEAN) COVID-19 Survey Archive has documented 170 studies, 56 datasets, 43 questionnaires, 318 analytical documents, and 1,859 survey questions related to SARS-CoV-2 (https://covid-19.parc.us.com/). Comparing responses across surveys will enable examination of potential biases of this Facebook based sample as well as those of other data collection efforts.

Frauke Kreuter

Appendix

Affiliations of authors of main paper

Frauke Kreuter University of Maryland, University of Mannheim, and IAB

Neta Barkay Facebook

Alyssa Bilinski Harvard University

Adrianne Bradford University of Maryland

Samantha Chiu University of Maryland

Roee Eliat Facebook

Junchuan Fan University of Maryland

Tal Galili Facebook

Daniel Haimovich Facebook

Brian Kim University of Maryland

Sarah LaRocca Facebook

Yao Li University of Maryland

Katherine Morris Facebook

Stanley Presser University of Maryland

Joshua A. Salomon Stanford University

Tal Sarig Facebook

Kathleen Stewart University of Maryland

Elizabeth A. Stuart Johns Hopkins Bloomberg School of Public Health

Ryan Tibshirani Carnegie Mellon University