Social surveys during the COVID-19 pandemic

Social Surveys During the COVID-19 Pandemic

Dr Roxanne Connelly and Professor Vernon Gayle

University of Edinburgh

The unfolding COVID-19 pandemic is resulting in unforeseen social and economic disruption, globally and locally. The rapid and unprecedented nature of the crisis requires the collection of contemporaneous research data. The disruptive nature of the pandemic produces practical challenges and methodological dilemmas for collecting suitable social science research data. In this chapter, we outline the problems, issues and opportunities associated with collecting research data using social surveys in this time of crisis.

There is a long history of using social surveys in social science research dating back to Seebohm Rowntree's Studies of York and Charles Booth's surveys of life and labour in London (Linsley and Linsley 1993). The social survey is best considered as a methodological approach, rather than a single technique. Historically, questionnaires have been the main data collection instrument used in social surveys, but interviews are frequently used to collect data. Increasingly a mixture of different modes of data collection are used, especially involving computers and new technologies (see De Leeuw 2005).

Social surveys are designed to study statistical populations. Statistical populations are aggregates of specific entities or cases. A statistical population commonly contains too many cases to expediently study, and it is therefore more practicable to draw a sample (i.e. sub-set) of cases. Integral to the survey method is the collection and analysis of data from a sample of a larger statistical population. The social survey method is attractive because collecting data from a sample is more practical, and sample data have statistically efficient properties.

A social survey is a methodology that generates a matrix of research data. The usual components of the matrix are cases and variables. Variables are the result of collecting systematic measurements, and cases are the entities under investigation. Cases will commonly be individuals, but they could be households, families, businesses, schools etc. Integral to the survey method is the organised collection of systematic measures across a set of cases to provide comparable data.

A fundamental aspect of constructing a survey sample from a statistical population is how well it represents, or reflects, the aspect of the statistical population under investigation. The techniques used to choose a sample can be divided into two broad classes, probability samples and non-probability samples.

A probability sample uses a formal statistical method to select cases from the statistical population for inclusion in the sample. In a probability sample, every case in the statistical population has a non-zero chance of being included in the sample, and cases within the sample will appropriately reflect the characteristics of the target population. Analysis of the sample data will therefore support inferences to the target population. A notable early example of probability sampling in social research is Arthur Lyon Bowley's New Survey of London Life and Labour (Kotz and Dale 2011) .

A probability sample requires a sampling frame, i.e. a list of cases within the target statistical population from which the sample can be drawn. The sampling frame might be an unrelated administrative resource, for example the Postcode Address File (PAF), a database containing all known postal delivery points in the UK.

Fundamental to choosing a probabilistic sample is the use of a statistically informed method of random selection. A conceptually straightforward approach is the simple random sample in which each case has an 'equal and non-zero' chance of inclusion, and where statistically random selection greatly reduces biases. Simple random samples generate cases that are representative of the target statistical population and will therefore support inferences to the target population. Furthermore, since Bernoulli posited the early version of the law of large numbers, it has been known that larger samples will tend to be more representative (Gigerenzer et al 1990).

In practice, large-scale social surveys usually adopt more sophisticated representative sampling approaches, for example stratification and clustering. The use of stratification explicitly addresses the problem of sub-populations within the overall target population varying, by independently sampling sub-populations. The adoption of clustering has practical advantages relating to collecting data, reducing the time and costs associated with undertaking face-to-face interviews. Social surveys with more sophisticated designs and sampling strategies have methodological benefits, but data analysis is more complex, requiring non-standard statistical analytical techniques.

Non-probability samples do not use a formal statistical method to select cases from the population for inclusion in the sample. In non-probability samples, every case in the statistical population does not have a non-zero chance of being included in the sample, and therefore the sample may not appropriately reflect the characteristics of the target population. There are various techniques for selecting non-probability samples. These include convenience sampling (alternatively known as haphazard or accidental sampling) where cases are chosen for relative ease of access; snowball sampling where a respondent refers, or makes a connection with, other potential respondents; and purposive sampling where respondents are specifically chosen because they are considered appropriate for the study. Quota sampling is a further method of non-probability sampling. It combines an element of probability sampling, by segmenting the target population into strata prior to establishing quotas, usually based on observable characteristics such as sex, age or ethnicity. Cases are then selected to fill quotas using non-probability sampling.

Overall, non-probability techniques do not use formal statistical methods to provide samples representative of target populations. A famous illustration of the deficiency of non-probability samples is the Literary Digest Poll for the 1936 United States Presidential Election (Squire 1988). The poll solicited 2.27 million individual responses, and predicted that Alfred Landon would win by a landslide, but he was beaten by Franklin D. Roosevelt. The poll did not represent the views of all voters because Literary Digest subscriber were notably more affluent than the general US population. This example also illustrates that having a very large sample size does not necessarily compensate for the unrepresentativeness of selection.

A more recent example, (2013) collected data for the Great British Class Survey (GBCS) using the BBC website and television, radio, and newspaper advertising. This resulted in a large sample size, however it was predominantly drawn from 'well-educated social groups' who

would typically complete surveys on the BBC website, and did not accurately represent the characteristics of the wider British population. Whilst the GBCS provided a wealth of contemporary data, the original GBCS sample survey data does not support reliable analyses of social difference and inference to the target population.

Having considered some general issues associated with the social survey method, we now turn to social science data collection and the COVID-19 crisis. The societal conditions that have rapidly emerged as a result of the pandemic present a series of general constraints for the collection of social and economic data, and some specific challenges for undertaking social surveys. Most notably, the need to protect people and minimise the spread of the COVID-19 virus from one person to another, has required new modes of safe human comportment. Face-to-face interviews are a survey data collection mode that has been proven to deliver high quality data (Groves et al 2011). Such activities were prohibited during the early stages of the pandemic. Data collection by telephone or on-line provide practicable alternatives but their relative advantages and disadvantages must be weighed (for a discussion see De Leeuw et al 2008).

Online questionnaires can be a valuable approach to collecting information (see Deutskens 2006, Dillman 2011, Stanton 1998). During the pandemic there has been a burgeoning trend in the distribution of online questionnaires, especially to collect information on social experiences such as living under lockdown. The data collection activities undertaken during the pandemic are frequently advertised via e-mail and on social media platforms. These surveys should be regarded as having been produced using non-probability (i.e. convenience) sampling, and data collection activities that encourage respondents to distribute questionnaires within their networks are using snowball sampling.

The demographics of social media users are known to differ from the general population (Mellon and Prosser 2017). The social and economic disruption that has flowed from the pandemic, and the resulting policy measures such as lockdown, the closure of schools and furloughing workers, may have had profound effects on the availability of survey respondents. For example, we envisage that furloughed workers are more likely to have time to respond to surveys distributed on social media compared with essential workers and parents that are caring for young children or providing home schooling.

Studies that have collected online questionnaire data using non-probability samples might be enhanced by using *post hoc* approaches to address issues of unrepresentativeness. A potential technique for this is propensity scoring adjustment (Schonlau et al 2004). This technique requires a separate reference data set based on a probability sample of the target population. The non-probability sample data set is combined with the reference data set, and a variable which indicates the sample origins of the two datasets is constructed. A statistical model (e.g. a logistic regression model) of sample membership is then estimated using a set of research and demographic indicators as explanatory variables. Predictions from the statistical model are then used to construct scores to reweight the non-probability sample to render it more reflective of the characteristics of the (probability selected) reference sample (Schonlau et al 2004).

The main obstacle to using propensity scoring adjustment is that it requires a reference sample which uses a probability sampling strategy, and the datasets will need to include a range of auxiliary variables that can be used in the model estimation required for the weighting process (Mercer et al 2018). In practice, a suitable data resource that appropriately represents a target

population during the pandemic is unlikely to be available. Furthermore, weighting techniques have been found to only marginally improve the estimates from non-probability samples, and routinely fail to address substantial biases (Mercer et al 2018, Tourangeau et al 2013, Vehovar et al 1999). Therefore, such *post hoc* adjustments to non-probability samples are not an especially feasible solution to the methodological challenge of improving questionnaire data using non-probability samples during the pandemic.

An elegant practical solution is for COVID-19 research to be piggybacked on existing infrastructural survey data resources. Several of the UK's large-scale social surveys have undertaken bespoke COVID-19 data collection exercises. This has allowed them to harness the benefits of existing representative large-scale samples. In addition, they have been able to utilize the expertise of their survey teams and to operationalize existing high-quality data collection procedures. This has resulted in contemporaneous COVID-19 data being rapidly made available to the social science research community.

Understanding Society, the UK Household Longitudinal Study (UKHLS) has blazed a trail in collecting COVID-19 data from a very large nationally representative UK sample. The UKHLS is one of the largest household panel surveys in the world and began by interviewing approximately 100,000 individuals from 40,000 households and it has made repeated contacts with then for a decade (Buck and McFall 2011). Monthly COVID-19 surveys have been collected via the internet since May 2020. In addition, telephone interviews will be used to contact sample members who do not use the internet. The UKHLS COVID-19 surveys collect data on the welfare of individuals, their families and their wider communities. The UKHLS will collect regularly repeated measures in order to support comparative temporal analysis, and will also include new measures as the pandemic unfolds. The data are being fast-tracked and made available to researchers shortly after collection via the UK Data Service.

COVID-19 survey data collection exercises are also being undertaken within the British Birth Cohort Studies. These are a series of well-established longitudinal studies which have continued to follow babies born in 1946, 1958, 1970 and 2000-02 to present day (see Pearson 2016). There are currently plans to collect surveys in May 2020 and August 2020. These surveys cover topics such as physical and mental health, time use, financial situations, employment, education and social connectedness. The English Longitudinal Study of Ageing (ELSA) is a study of older citizens (originally aged 50 years and over) (Steptoe et al 2013). ELSA will collect COVID-19 data in May 2020 and September 2020 using a combination of internet and telephone interviews. The measures will focus on issues such as physical and mental health, financial circumstances, social connectedness and health behaviours. These large-scale data resources will provide invaluable information which will support high quality research across a range of social science disciplines, and ultimately contribute to the evidence base and to policy formulation and planning.

In conclusion, the severity of the COVID-19 virus and its impact on social and economic life is difficult to overemphasize. It is critical that appropriate data are collected that can make demonstrable contributions to understanding changes in societal conditions, and ultimately enabling social research that benefits individuals, organizations and wider society. High quality valid and reliable data are the *sine qua non* of research excellence.

Undertaking social surveys using the internet and social media with non-probability samples, may initially seem attractive because of perceived needs for rapid data collection. This method

of sampling and selection compromises survey data quality. Drawing reliable inferences about target populations from sample data with non-probability samples is a serious and insoluble problem that must not be overlooked (Schonlau et al 2002, Schonlau et al 2004).

The global nature and the ubiquitous consequences of the pandemic have made it an all-encompassing news item for most of 2020. An exceptional feature of the pandemic period is that UK media outlets have reported an unparalleled amount of data and statistical information, especially in the form of data visualizations. The extent to which media reports have suitably considered how social science data are collected, and the implications that this has for producing reliable and valid results, is not readily comprehensible.

Concomitantly, a virtual army of amateur (or armchair) statisticians have emerged, using social media to present and share analyses. Sampling is an esoteric aspect of the survey method. We have concerns that non-professional researchers may not always appreciate the scope and limitations of different sampling and selection methods and therefore cannot appropriately assess the reliability and validity of subsequent results.

We advocate for greater research transparency because it increases the capacity to understand how the research was conducted, helps other scholars evaluate analyses, aids the detection of errors and inconsistencies, facilitates the incremental development of work, contributes to limiting negative research practices, provides extra safeguards against nefarious practices, and improves confidence in results (Connelly et al 2020). The specific details of sampling and selection methods are often buried in supplementary materials associated with research projects. We contend that such information should be highlighted and made more easily accessible because it has important consequences for assessing the validity and reliability of research findings.

Collecting social science data relating to the COVID-19 crisis as part of existing large-scale data collection enterprises is an ingenious, practicable solution to the problem of gathering suitable, high-quality data. Locating this contemporaneous data within existing longitudinal data series provides unique opportunities to understand the specific impact of the COVID-19 crisis and social and economic change.

References

Buck, N.andMcFall, S. (2011) 'Understanding Society: design overview', *Longitudinal and Life Course Studies,* 3(1): 5-17.

Connelly, R., Gayle, V.andPlayford, C. (2020) 'Undertaking Transparent and Reproducible Data Analysis', *The Sage Dictionary of Social Research Methods*, London, SAGE Publications.

De Leeuw, E. D. (2005) 'To mix or not to mix data collection modes in surveys', *Journal of Official Statistics,* 21(5): 233-55.

De Leeuw, E. D., Hox, J. J.andDillman, D. A. (2008) *International handbook of survey methodology,* New York: Taylor & Francis Group.

Deutskens, E. C. (2006) 'From paper-and-pencil to screen-and-keyboard: studies on the effectiveness of internet-based marketing research'.

Dillman, D. A. (2011) *Mail and Internet surveys: The tailored design method--2007 Update with new Internet, visual, and mixed-mode guide,* New York: John Wiley & Sons.

Gigerenzer, G., Swijtink, Z., Porter, T.andDaston, L. (1990) *The empire of chance: How probability changed science and everyday life,* Cambridge: Cambridge University Press.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E.andTourangeau, R. (2011) *Survey Methodology,* Hoboken, NJ: John Wiley & Sons.

Kotz, S.andDale, A. I. (2011) *Arthur L Bowley: A Pioneer in Modern Statistics and Economics*: World Scientific.

Linsley, C. A.andLinsley, C. L. (1993) 'Booth, Rowntree, and Llewelyn Smith: a reassessment of interwar poverty 1', *The Economic History Review,* 46(1): 88-104.

Mellon, J.andProsser, C. (2017) 'Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users', *Research & Politics,* 4(3): 2053168017720008.

Mercer, A., Lau, A.andKennedy, C. (2018) *For Weighting Online Opt-In Samples, What Matters Most?,* Washington, DC.: Pew Research Center.

Pearson, H. (2016) *The life project: The extraordinary story of 70,000 ordinary lives*: Catapult.

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., Le Roux, B., Friedman, S.andMiles, A. (2013) 'A new model of social class? Findings from the BBC's Great British Class Survey experiment', *Sociology,* 47(2): 219-50.

Schonlau, M., Ronald Jr, D.andElliott, M. N. (2002) *Conducting research surveys via e-mail and the web,* New York: Rand Corporation.

Schonlau, M., Van Soest, A., Kapteyn, A., Couper, M.andWinter, M. (2004), Adjusting for selection bias in Web surveys using propensity scores: the case of the Health and Retirement Study. *Proceedings of the section on survey statistics, American Statistical Association*. pp. 4326-33.

Squire, P. (1988) 'Why the 1936 Literary Digest poll failed', *Public Opinion Quarterly,* 52(1): 125-33.

Stanton, J. M. (1998) 'An empirical assessment of data collection using the Internet', *Personnel psychology,* 51(3): 709-25.

Steptoe, A., Breeze, E., Banks, J.andNazroo, J. (2013) 'Cohort profile: the English longitudinal study of ageing', *International journal of epidemiology,* 42(6): 1640-48.

Tourangeau, R., Conrad, F. G.andCouper, M. P. (2013) *The science of web surveys,* Oxford: Oxford University Press.

Vehovar, V., Manfreda, K. L.andBatagelj, Z. (1999), Web surveys: Can the weighting solve the problem. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. pp. 962-67.