Theme 1

Issues of Coverage and Sampling in Web Surveys for the General Population: An Overview

By Peter Lynn, University of Essex, plynn@essex.ac.uk

Synthesis paper for NCRM Web Survey Network opening conference

1. Constraints and Issues

A major barrier to the use of web data collection methods for probability-based surveys of the general population is the difficulty of achieving good population coverage. This paper provides an overview of the issues to be faced and of possible solutions, some of which are as yet untested.

The nature of the issues to be faced, and the possible solutions, are dependent on the main motivation for wanting to consider web data collection, which can be related either to cost reduction or to improving coverage and participation. We therefore outline the implications of each type of motivation. We also emphasise that the issues and solutions are somewhat different for longitudinal surveys and cross-sectional surveys. And there is an important distinction between single-mode web surveys and mixed-mode surveys that include web as one of the modes. These two different types of surveys present rather different challenges.

Motivations for Data Collection by Web

Replacing interviewer-administered methods of survey data collection with web data collection has the potential to deliver considerable cost savings, as the marginal cost of each extra completed web questionnaire is very much less than the marginal cost of each extra interview conducted by an interviewer. Web data collection also has the potential to improve sample coverage and participation, as some types of people who tend to be under-represented in interviewer-administered surveys may be more likely to participate by web. However, these different objectives – cost reduction and improved coverage and participation – appear to require very different solutions and may therefore be incompatible (Lynn, 2013). Improved coverage and participation requires that all sample members who would have taken part in an interviewer-administered survey continue to take part and that, in addition, some people who would not have taken part in an interviewer-administered survey should take part by web.

NCRM Web Survey Network partner organisations:

NatCen Social Research · CLS · iSER Institute for Social & Economic Research · UNIL | Université de Lausanne · NCRM National Centre for Research Methods

Longitudinal vs. Cross-sectional Surveys

Longitudinal surveys allow the possibility of collecting information from sample members that can help to determine the mode(s) in which they are willing and able to participate and also information that can facilitate making contact in different modes (specifically, email addresses). Thus, there may be less imperative with a longitudinal survey to find ways of sampling that will permit immediate data collection by web. Instead, the first wave of the survey may be carried out by traditional means while still, over the life of the survey, shifting a large part of the data collection to web.

Single-Mode vs. Mixed-Mode Surveys

The decision as to whether a survey should collect all data in a single mode or should employ a mix of modes is a complex one that should reflect consideration of many factors. With respect to sampling and coverage a single-mode web survey is likely to prove more challenging than a mixed mode survey. It may be necessary to accept substantial compromises. While mixed-mode approaches may be undesirable in terms of issues related to measurement and complexity, they may be able to offer solutions to coverage problems that cannot easily be solved in the context of a single-mode web survey.

2. Coverage issues with single-mode web

If the intention is to carry out a single-mode web survey of the general population, two broad classes of approach are possible. One is to restrict coverage of the survey to web users. The other is to find a way of having non web-users respond by web. We discuss here the under-coverage implications of the first approach and the practical implications of the second approach.

Implications of Restricting Surveys to Web Users

If no attempt is made to include people who are not web users, survey samples will be biased in important ways (see the presentation by Callegaro at this conference). Non-users of the web are quite distinct in terms of socio-demographic and other characteristics. However, they are less distinct than was the case several years ago and they are also decreasing in proportion, so under-coverage bias may conceivably become less of an issue in the future. Nevertheless, to claim that a sample of web users can provide adequate coverage of the total population, even after sophisticated statistical adjustment and estimation procedures have been applied, currently remains a strong claim indeed.

In assessing the potential problem of under-coverage, care should be taken with the concept of a web user. If, for example, we mail a random sample of the population requesting participation in a web survey, it is the overall nature of non-response bias in which we are interested, a large part of which may be caused by what we think of as under-coverage of the web. In other words, some sample members will not have access to the web or will not be able to use it. But measures of whether there is internet access in the home and whether a person is a regular or occasional web user will provide only an approximation to the group who are actually able to respond to a web survey. Some people without internet access at home may have access at their place of work or education, at someone else's home, or in a

public place such as a library or internet café. And increasingly access is wireless, so people can have access almost anywhere regardless of whether they have it at home. And some people who do not define themselves as web users may nevertheless take part in a web survey, perhaps with help from a partner, family member or friend, or indeed via a proxy response.

Various statistical adjustment procedures have been proposed as a means of overcoming the undercoverage bias inherent in web-only samples (e.g. Bethlehem 2010, Duffy et al 2005, Lee & Valliant 2009, Schonlau et al 2004, Schonlau et al 2009, Valliant & Dever 2011). These are routinely implemented, for example by proponents of online access panels. The procedures typically involve weighting through a pseudo-randomisation approach in which the probability of a person participating in the web survey is estimated via a propensity scoring approach, often followed by calibration to some known population characteristics. Alternatively, model-based estimation of outcomes of interest can be used (Deville 1991). However, any procedure of this kind relies on the assumption that available information about differences between the online population and the total population explains, in a statistical sense, differences in the target variables or estimates. In other words, to be effective the procedures require that data are missing at random (Little & Rubin, 2002).

Implications of including non- users

An approach to achieving full population coverage is to supply hardware, internet access and training to sample members who do not already have these things. This is only cost-effective if each sample member takes part repeatedly, so it is only done when setting up a panel, not for an individual survey. The earliest examples of this are from the UK (Clemens 1984) and the *Telepanel* in the Netherlands, which began data collection in 1986 (Saris 1998). At that time hardly any households possessed a computer that was suitable for survey completion or had internet access. Consequently, these early approaches involved supplying all sample households with the same equipment. The novelty of having such equipment in the home may have been a contributing factor to the enthusiasm with which households apparently took part in those surveys.

More recently, some surveys have adopted a strategy of relying on sample members' own computing equipment wherever possible but providing equipment to the remaining minority of households who would not otherwise be able to participate. Each of these surveys provides households with some form of simple computer and internet access. The LISS panel[1] in the Netherlands, for which recruitment started in 2007, provides broadband access and a small, simple PC called the *SimPC* (see the presentation by Scherpenzeel at this conference). The KnowledgePanel[2], run by Knowledge Networks in USA and for which recruitment began in 2009, supplies households with a netbook and internet service. The German Internet Panel[3], recruiting in 2012, also provides a simple personal computer, the *Ben PC*, and internet access.

[1] http://www.lissdata.nl
[2] http://www.knowledgenetworks.com/knpanel/
[3] http://reforms.uni-mannheim.de/english/internet_panel/home/

Meanwhile, the methodology of the new French probability-based internet panel, ELIPSS, is more like a modern-day equivalent of the Dutch telepanel. Sample members recruited to ELIPSS will each be issued with a *Samsung Galaxy Tab 7* touch-screen tablet and given a 3G subscription, so all respondents will be using identical technology. This approach at least ensures that all respondents should see the survey questions identically, with no differences in visual appearance caused by variations in browser, screen size or settings.

3. Coverage issues with mixed-mode

Another approach to ensuring the inclusion of households without web access and a suitable computer is to allow such households to participate in the survey by other means, such as a mail questionnaire or a telephone interview. This therefore results in a mixed mode design, where as many households as possible participate by web but the residual use an alternative mode (e.g. Dillman & Messer 2010, Messer & Dillman 2010, Millar & Dillman 2011). An example of such a design is the Gallup Panel in the USA, in which about 70% of sample households participate in web surveys while the remainder complete mail surveys, the initial recruitment having been done entirely by telephone (Rao et al 2010, Rookey et al 2008). Similarly, longitudinal surveys which begin as a single-mode interviewer-administered survey can attempt to convert participants to responding by web, resulting in a situation where some sample members respond in a mix of modes (over time) while others continue to respond solely in an interviewer-administered mode.

Such mixed mode designs have the advantage that cost savings are maximised by having as high a proportion as possible of sample members complete by web, while under-coverage due to lack of web access is avoided. But there are also potential disadvantages. The first is the risk of differential measurement error between the modes, which affects any mixed mode survey (see the presentation by Calderwood at this conference). Additionally, field management becomes complex, the speed advantages of web data collection are lost, and it becomes difficult to ensure that data from each mode are collected at the same time (and therefore refer to the same time periods, introducing analysis complexity).

4. Sampling issues with single-mode web

There are no sampling frames of the general population that include email addresses. This makes it impossible for the first approach to be made by email with a web-link. Instead, single mode web surveys of the general population must begin with an approach by another mode involving a request for the sample member to go online and access the survey. The cheapest way of making this first approach is by mail. In countries with some kind of population register that can be used as a sampling frame, a mail approach is relatively unproblematic, but this is not the case in the UK. Probability-based general population surveys in the UK tend to rely on sampling frames of addresses, notably the Postcode Address File (PAF). PAF does not include any names of residents at the addresses, nor any indicator of how many individuals reside at each address. Thus, the identification and selection of individuals to take part in a survey must be undertaken as part of the survey field work process. In face-to-face interview surveys the interviewer can control this process, but with a mail approach it is necessary to rely on the unnamed recipient of the mail diligently following instructions to identify which resident(s) should complete the survey. The instructions can be in the mailing letter or, to reduce the potential ability of the respondent to

control the selection process, the selection procedure can form the first part of the online instrument. In either case, such procedures are error-prone (Battaglia et al 2008, Cabinet Office 2012). In many cases the wrong person will complete the survey, either due to misunderstanding or due to deliberate choice on the part of the recipient of the mailing.

To have better control over the respondent selection process would require interviewer administration of the initial approach. This is of course possible but immediately undermines the cost advantages of a web-only survey.

Other approaches to sampling for general population web surveys do not attempt to retain a probability basis for selection. Respondents are recruited through pop-up surveys, online advertisements or through direct emails sent to samples selected from lists constructed from various sources. These are generally referred to as opt-in panels. Such approaches are much less costly than attempting to recruit a probability sample, but they require an alternative inferential paradigm (Bethlehem 2010). Standard statistical sampling theory cannot be applied and it cannot be assumed that the sample reflects the characteristics of the population in important respects. Instead, inference must involve some form of statistical model of the relationship between the sample and the population. Various approaches to model-based inference are used by opt-in and similar panels. The extent to which these approaches are fit for purpose depends on the accuracy of the underlying assumptions, which in turn often depends on whether it is possible to test the assumptions empirically and modify if necessary.

5. Sampling issues with mixed-mode

Surveys that attempt complete population coverage through the use of a mixed mode approach can take advantage of partial frame information about contact details required for different modes of approach. For example, a frame with good coverage and complete address information, but telephone numbers for only a proportion of frame units can be used for a mixed telephone/ face-to-face approach where as many sample members as possible are approached by telephone with a face-to-face visit for the remainder. This idea could be extended to frames that contain email addresses for only a proportion of individuals, though such frames do not yet exist for the UK general population.
Instead, at present it is necessary to make the initial approach in a single mode for all sample members. The least costly mode of approach is mail. To maximise the proportion of sample members who respond by web, an efficient design may be to provide only the web option in the initial mailing and reminder(s) (Millar & Dillman 2011). However, a subsequent follow-up by a different mode is then required to those who have not responded by web. To further reduce costs, it is tempting to include a paper self-completion version of the questionnaire in the initial mailing. However, such simultaneous mixed-mode designs, in which the sample member is faced with a choice of modes in which to respond, tend to result in lower overall participation rates than sequential designs in which the sample member is only offered one mode on each occasion (Millar & Dillman 2011, Smyth et al 2010). Hybrid approaches are also possible, such as offering only a web option in the initial mailing but inviting the sample member to request a paper questionnaire if they would prefer that, or offering only a web option initially but indicating that the respondent will later be given an opportunity to participate in a different mode if web does not suit them. The relative merits of these different elements of mixed mode designs have yet to be fully evaluated.

However, in the UK even these designs that involve initially mailing all sample members are constrained by the absence – referred to above – of a high-quality sampling frame that includes individual names. It may be preferred instead to adapt the mode of initial approach to the available sampling frame data. As good frames are available of telephone numbers and addresses, the first approach could be made either by phone or face-to-face. However, the considerable expense of a face-to-face approach means that this is only likely to be considered for longitudinal surveys, in which the cost of the initial (recruitment) contact can be defrayed over several subsequent online surveys/waves.

6. Conclusion and Discussion

Currently, the limited options for sampling frames and sampling methods in the UK severely constrain the use of web in probability-based general population surveys. Good population coverage can be achieved through use of sampling frames traditionally used for interviewer-administered surveys, but unless the initial approach is then interviewer-administered, the researcher loses control over respondent selection. Web data collection is mainly considered only for longitudinal surveys, in which respondent names are known and email addresses can be collected at previous waves. For one-time surveys a mail invitation to a web survey can be used, but errors of respondent identification are likely and are hard to detect. The prospects for web data collection from the general population would change dramatically if a population register were to become available and especially if this were to include email addresses. However, this seems highly unlikely, at least in the foreseeable future. For these reasons, non-probability methods are likely to remain attractive to many for some time to come.

References

Battaglia, M.P., Link, M.W., Frankel, M.R., Osborn, L. and Mokdad, A.H. (2008). An evaluation of respondent selection methods for household mail surveys. *Public Opinion Quarterly* 72 (3): 459-469.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review* 78(2):161-88.

Deville, J.-C. (1991). A theory of quota surveys. *Survey Methodology* 17:163-81.

Dillman, D.A. & Messer, B.L. (2010). Mixed mode survey. Chapter 17, 551-574, in P. Marsden and J. Wright (eds.) *Handbook of Survey Methodology*. Emerald Publishing: Bingley.

Duffy, B., Smith, K., Terhanian, G. and Bremer, J. (2005) Comparing data from online and face-to-face surveys, *International Journal of Market Research* 47(6): 615-639.

Jäckle, A., Lynn, P. and Burton, J. (2012) Going online with a face-to-face household panel: Initial results from an experiment on the UK Household Longitudinal Study Innovation Panel. Paper presented at the annual conference of the *Social Research Association*, London, December.

Little R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd Edition), John Wiley, New York.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods and Research* 37(3): 319-43.

Lynn P. (2013) Mixing modes on household panel surveys: opportunities, constraints and challenges. Paper presented at the 7th *International Conference of Panel Data Users*, Lausanne, February.

Messer, B.L. and Dillman, D.A. (2010). Using address-based sampling to survey the general public by mail vs. "web plus mail". *Social andEconomic Sciences Research Center Technical Report* 10-13. Washington State University: Pullman.

Millar, M.M. and Dillman, D.A. (2011). Improving response to web and mixed-mode surveys. *Public Opinion Quarterly* 75 (2): 249-269.

Rao, K., Kaminska, O. and McCutcheon, A.L. (2010) Recruiting probability samples for a multi-mode research panel with internet and mail components. *Public Opinion Quarterly* 74 (1): 68-84.

Rookey, B.D., Dillman, D.A. and Hanway, S. (2008). Does the inclusion of mail and web alternatives in a probability-based household panel improve the accuracy of results? Paper presented at the *American Association of Public Opinion Research Conference*, New Orleans.

Saris, W. (1998). Ten years of interviewing without interviewers: the telepanel. Chapter 21, 409-429, in M.P. Couper et al (ed.s), *Computer Assisted Survey Information Collection*, John Wiley, New York.

Schonlau, M., Zapert, K., Simon, L.P., Haynes Sanstad, K., Marcus, S.M., Adams, J., Spranca, M., Kan, H., Turner, R. and Berry, S.H. (2004) A comparison between responses from a propensity-weighted web survey and an identical RDD survey, *Social Science Computer Review* 22: 128-138.

Schonlau, M., van Soest, A., Kapteyn, A. and Couper, M. (2009) Selection bias in web surveys and the use of propensity scores, *Sociological Methods and Research* 37(3): 291-318.

Scott, A., Jeon, S.H., Joyce, C.M., Humphreys, J.S., Kalb, G, Witt, J. and Leahy, J. (2011) A randomised trial and economic evaluation of the effect of response mode on response rate, response bias, and item nonresponse in a survey of doctors, *BMC Medical Research Methodology* 11:126.

Smyth, J.D., Dillman, D,A,, Melani Christian, L. and O'Neill, A. (2010). Using the internet to survey small towns and communities: limitations and possibilities in the early 21st century. *American Behavioral Scientist* 53:1423–48.

Valliant, R. and Dever, J.A. (2011) Estimating propensity adjustments for volunteer web surveys, *Sociological Methods and Research* 40(1): 105-137.