

Web surveys: methodological problems and research perspectives¹

Silvia Biffignandi
Bergamo University,
Via Caniana n. 2, 24127 Bergamo(It)
silvia.biffignandi@unibg.it

Jelke Bethlehem
Statistics Netherlands
P.O. Box 24500, 2490 HA The Hague
jbtm@cbs.nl

Abstract: The paper presents a framework of the current problems and related literature on Internet/Web surveys. Attention is focused on the classification of topics with reference to the different research contexts (official statistics, academic research, market research), and on the identification of critical research, open questions and trends.

Keywords: self-selection, inference, access panel, official statistics, market research.

1. Introduction

There is a growing interest in using web surveys (also called Internet surveys or online surveys) for data collection. This is not surprising as the Internet provides easy access to a large group of potential respondents, and conducting such a survey is relatively cheap and fast. Moreover, costs do not depend too much on the number of interviews. Notwithstanding the appeal of web surveys, there are serious methodological problems. These problems have to be solved in order to obtain reliable survey outcomes. In practice, many surveys are carried out without paying much attention to the problems involved in each step of the survey process and on the reliability of the collected data. It is therefore very important: a) to draw the attention of the practitioners of web surveys to methodological problems and to the need to be cautious when carrying out such surveys and using the results for decision support; and b) to pay attention to studies investigating methodological aspects that may undermine data quality of web surveys. In our paper, we present a short review of the main methodological problems of web surveys, the literature on these problems and research trends. WebSm (WebSurvey Methodology) is a European project that has created – since 2003 - a network of partners² which have set up a website which publishes in an well organized framework the literature, software references and comments (Lozar Manfreda, Vehovar, 2006). The research team formed for this project is currently studying methodological problems in web surveys. This research group periodically (each two years) organizes a workshop discussing new frontiers and progress in this research area. Already four workshops

¹ The present paper is financially supported from the ex 60% research funds 2007 and 2008 of the University of Bergamo, research responsible Silvia Biffignandi.

²Partners of the European project WebSM (Web Survey Methodology) are: University of Ljubliana, University of Bergamo, Linköping University, ZUMA.

have been organized; the last one at the University of Bergamo in 2009. There are two major research areas that are of great importance:

1) design of survey questionnaires. Web surveys can use all kinds of new tools such as sound, video, colours and animation. How such tools will affect response behaviour is not yet completely clear. Therefore, one should be careful when designing a web survey;

2) recruitment of participants. How can a representative sample be selected? If a sample is not representative, can it be corrected such that it becomes representative? And what are the effects of nonresponse?

The impact of these problems and possible solutions depend to a large extent on the type of population which is studied (for example, approaches may be different for household surveys and establishment surveys) and on the type of organization carrying out the web survey (official statistics, universities or market research companies).

It is important to use a proper terminology. Internet surveys are surveys that collect data over the Internet. This includes e-mail surveys, where the respondent returns the questionnaire by email, attaching responses written in a word, excel document, or other software which is available to the respondent (for example Blaise). So, the questionnaire is completed off-line. Web surveys are completed on-line. The questionnaire is presented to respondents as a set of web pages. Answers to questions are transmitted immediately after clicking on a submit / next button.

2. Questionnaire design

Many methodological problems are common to both internet and web surveys. It should be noted that questionnaire design is more critical and relevant for web surveys since they may use special tools such as visual animation, sounds and colours. Therefore, the focus of this paper is on web surveys, but it should be taken into account that many problems are shared with internet surveys.

Web surveys are self-administered. Therefore, user-friendliness is of the utmost importance. If respondents encounter difficulties when completing the questionnaire, they quit, leaving the researcher with incomplete and often unusable data. Visual elements, pictures, colours and sounds are tools that can make the questionnaire more attractive. Technical limitations (for example low modem speeds), however, may prohibit their use, and an overload of not always relevant information may distract respondents. The impact of advanced graphical techniques and sound tools on the respondent's behaviour needs more study. It has to be demonstrated that they have a positive impact on the survey participation and data quality. An extended discussion and review of the features of web questionnaire design is given in Couper (2008). Recent studies have underlined that web questionnaire design problems should be treated in a manner consistent with the way in which sophisticated questionnaire designs problems are addressed (Dillman, 2007). Questionnaire length is an important aspect both with respect to the data collection mode and with respect to the target population of a web survey. Actually, in web surveys the questionnaire should be quite short (Czaja, Blair, 2005). Nevertheless if panels or closed group or special target population is investigated, a longer questionnaire could be effective as well.

3. Data collection modes, recruitment and inference

Web surveys are substantially affected by coverage, because not every member of a target population may have access to the internet. Moreover, it is often difficult to select a proper probability sample because a sampling frame is lacking.

More and more, a web survey is considered as one of the modes in a mixed-mode survey. If the objective is to keep survey costs as low as possible, one could start with a web survey, re-approach the nonrespondents by means of telephone interviewing, re-contact the nonrespondents in a face to face survey. Unfortunately, a mixed-mode survey introduces new problems, like mode effects, e.g. the same question is answered differently in a different mode of data collection. Occurrence and treatment of mixed-mode effects need further investigation.

In principle, web surveys can be carried out very quickly. It turns out response times to internet or web surveys are different from other modes and best procedures have to be identified. For instance, Biffignandi et al. (2000 and 2002) have studied response behaviour and the impact of the frequency of reminder procedures. Recent studies have been focusing on the best reminder tools, including Sms (which seems to perform well). Other studies concentrate on keeping respondents focused on the relevant parts of the computer screen, and keeping distraction to a minimum. One of the analysis techniques used is eye-tracking analysis.

Representativity of web surveys is an important issue. Scientifically meaningful results can only be obtained if proper probability samples are selected and the selection probabilities are known and positive for every member of the population. Unfortunately, many web surveys rely on self-selection of respondents. The survey is simply put on the web. Respondents are those people who happen to have internet, visit the website and decide to participate in the survey. The survey researcher is not in control of the selection process. These surveys are called self-selection surveys. See Bethlehem, Stoop (2007) for a discussion and Bethlehem (2008) for the bias caused by self-selection.

Similarly, Internet based panels (so-called access panels) are constructed by wide appeals on well-visited sites and Internet portals. So, they are also based on self-selection. At time of registration, basic demographic variables are recorded. A large database of potential respondents is created in this way. For future surveys, samples are selected from this database. Only panel members can participate in these web panel surveys. The target population is unclear. One can only say that it consists of people who have an Internet-connection, who have a non-zero probability of being confronted with the invitation, and decide to participate in the panel. Research in the Netherlands has shown that panel members differ from the general population (Vonk et al., 2006).

Access panels have the advantage that values of basic demographic variables are available for all participants. So the distribution of these variables in the survey can be compared with their distribution in the population. Over- or under-representation of specific groups can be corrected by some kind of weighting adjustment technique. However, there is no guarantee that this leads to unbiased estimates. To allow for unbiased estimation of the population distribution, a reference survey can be conducted that is based on a true probability sample from the entire target population. Such a reference survey can be small in terms of the number of questions asked. It can be limited to the so-called 'webographic' or 'psychographic' questions. Preferably, the sample size of the reference survey should be large to allow for precise estimation. A small sample size results in large standard errors of estimates (Bethlehem, 2007). Since most (but not all) access panels are based on self-selection, it is impossible to compute unbiased estimates of population characteristics. In order to properly apply statistical

inference, probability sampling in combination with a different data collection modes can be applied for panel recruitment. For example, a sample is selected from the general population and respondents are recruited using e.g. CAPI or CATI. Respondents without internet are provided with Internet access. See Scherpenzeel (2006) for an example. A probabilistic sampling design can be achieved by using some specific methods, such as random digit dialling (RDD). Some probability-based Internet panels have already been constructed in this way (for instance, Huggins and Krotki, 2001). Another approach to correct for lack of representativity is applying propensity scoring methodology. Propensity scores has been used to reweight web survey results (Schonlau et al., 2009, Biffignandi et al., 2003 and 2005).

References

- Bethlehem J. (2007) *Reducing the bias of Web Surveys Estimates*, CBS Discussion paper 07001, Voorbug/Heelen.
- Bethlehem J.G. (2008) *How accurate are self-selection web surveys?*. Discussion Paper 08014, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
- Bethlehem, J.G. , Stoop, I. (2007) Online panels - a Paradigm Theft? In M. Trotman et al. (Eds.), *The challenges of a changing world*. Southampton, UK: ASC, 113-131.
- Biffignandi S., Pratesi M. (2005) Indagini Web: propensity scores matching e inferenza. Un'analisi empirica e uno studio di simulazione, in *Integrazione di dati di fonti diverse*: Falorsi P., Pallara A. Russo A. (editors), F. Angeli, Milano, 131-158.
- Biffignandi S., Pratesi M., Toninelli D. (2003) Potentiality of propensity scores methods in weighting for Web surveys: a simulation study based on a statistical register, ISI Conference, Berlino, August (CD).
- Biffignandi S. , Pratesi M. (2002) Modeling The Respondents' Profile In A Web Survey On Firms In Italy, In A. Ferligoj, A. Mrvar (ed.), *Developments in Social Science Methodology, Metodoloski zvezki*, 18, 171-185.
- Biffignandi S., Pratesi M. (2000) *Modelling firms response and contact probabilities in Web surveys*, ICESII Proceeding Conference, Buffalo, USA, June 2000.
- Czaja R., Blair J. (2005) *Designing surveys: A guide to decisions and procedures* (second edition), Sage, Oaks CA.
- Couper M.P. (2008) *Designing Effective Web Surveys*, Cambridge Univ.Press, N.Y.
- Dillman D. (2007) *Mail and Internet Surveys, The Tailored Design Method*, Wiley .
- Huggins V., Krotki K. (2001) Implementation of nationally representative web-based surveys, *Proceedings of the Annual Meeting of the ASA*, August 5-9.
- Lozar Manfreda K., Vehovar V. (2006) Web survey methodology (WebSM) portal, in B. Aaron, D. Aiken, R.A. Reynolds, R. Woods, J.D. Baker (Eds), *Handbook on Research on Electronic Surveys and Measurement*, Hershey PA, 248-252.
- Scherpenzeel, A. (2008), An online panel as a platform for multi-disciplinary research. In: I. Stoop & M. Wittenberg (eds.), *Access Panels and Online Research, Panacea or Pitfall?*, Aksant, Amsterdam, 101-106.
- Schonlau M., van Soest A., Kapteyn A., Couper M. (2009) Selection Bias in Web Surveys and the Use of Propensity Scores, *Sociological Methods & Research*, Vol. 37, No. 3, 291-318.
- Vonk, T., Van Ossenbruggen, R. and Willems, P. (2006) The effects of panel recruitment and management on research results, a study among 19 online panels. *Panel Research 2006, ESOMAR World Research*, ESOMAR Publ.Serv. 317, 79-99.