

Modeling non-sampling errors and participation in Web surveys

Silvia Biffignandi

Abstract A new survey paradigm now takes non-sampling errors into account as part of the information regarding total error (TSE) in survey evaluation. In this context, the whole business process is analysed as a possible source of error, as well as being a quality indicator and a starting point for improvement. In Web surveys non-sampling errors are very important and need to be considered seriously. During the entire procedure, from the design of the survey to the end of the process, a number of errors might occur. There are two aspects typical for Web surveys: 1) statistical inference and bias, which tend to be more difficult to handle due to self-selection, to frame coverage troubles and non-response peculiarities; 2) with respect to other survey modes, a large amount of para-data can be collected during the survey process and this is potentially highly valuable in understanding how to improve surveys (if appropriately investigated). In this paper, the focus is on non-response and under-coverage errors and on survey participation, key issues with regard to non-sampling framework. Two approaches to dealing with these issues (bias oriented and behavioural oriented) are briefly presented together with a description of some of the research results.

Key words: sample design and estimation, Web survey, non response

1 Introduction

Survey methodology has quite recently undergone a paradigm shift. In the past the focus was on the effects of sampling errors on estimates. Now the focus has shifted to the causes of survey errors and how to prevent them. This new approach has extended the concept of error beyond the statistical inference framework and investigates

¹ Silvia Biffignandi, Bergamo University; email: silvia.biffignandi@unibg.it

Research grant: PRIN 2007, Pisa University local unit and BIFF60% local research project University of Bergamo, 2008.

theoretical and operational ways of reducing various types of error. It aims at also taking facts and behaviors occurring during the survey process into account. The old and new approaches are not in fact at odds with each other, rather they are complementary. To some extent, taking both approaches into accounts leads back to the concept of Total Survey Error (TSE). However, a key point is that the new paradigm should now stress the complementarization of the errors rather than integration in a complex model, which is quite an ambitious task. Thus, total survey error is concerned with errors that can occur in the whole survey process. It includes sampling error (the central error in the traditional approach) and non-sampling errors, related to the data collection and processing procedures. Total survey error is meant as the difference between the actual value in the population (Y) and survey estimates (Y^*), i.e. $Y - Y^*$. It includes sampling error component (the central error in the traditional old approach) and non-sampling error components, related to data collection and processing procedures. The latter components are highly emphasized by the new paradigm and it is stated that they could be numerically greater than the sampling error, traditionally used in survey result evaluation.

Non-sampling errors are different in nature; they relate to non-response, to overcoverage and undercoverage problems (frame errors), as well as to measurement and processing errors. Non-sampling errors can be grouped into two main types: systematic and random. Systematic errors (or bias) make survey results unrepresentative of the target population by distorting the survey estimates in one direction. Random error can distort the results in any given direction but tend to balance out on average.

Although in many cases we cannot directly calculate a numerical measure for them, and never include them in a general total survey error model, there are a number of approaches that may be used in order to control non-sampling error. Some approaches simply involve investigating the entire survey process and understanding what the sources of error are and where they occur in the survey. Other approaches are related to modeling specific survey situations.

Using a collaborative approach towards a compartmentalized vision of the total survey error, in my paper I focus on Web surveys and I analyze under-coverage and non-response issues from the systematic error point of view. Additionally, I describe some recent approaches to evaluating participation in Web surveys. Such approaches in some cases relate to behavioral and psychological theories.

2 Non-sampling errors in Web surveys

The spread of Web data-collection tools is causing attention to shift more and more to non-sampling error components. This is because Web surveys are self-administered, they often suffer from coverage problems (especially under coverage), and, since not all the target population has Web access and so cannot be part of the frame, non-respondents are not represented as part of the whole population. Also their profile with respect to the total population is often unknown. In addition, Web surveys on average have lower response rate than other survey modes.

Given the degree to which the Web survey context is affected by the above mentioned critical features, understanding non-response, coverage and participation behavior, rather than simply computing the response rate, becomes highly important and a challenge for strategies aimed at improving survey quality. Till now many efforts

and studies have been carried out to increase the total response rate as a way of improving survey quality. But, it should be taken into account that the overall response rate is a very simple measure of survey quality, even if currently used as a meaningful indicator. To some extent this measure could be useful in identifying weak points in the process (for instance a large refusal amount might be due to a bad-contact process), but it fails to consider that people who have web access, or that are respondent in a web survey, could significantly differ from other units. Moreover, overall response rates do not give information regarding the response propensity of different respondent subgroups (late respondents versus first respondents, and socio-demographically different subgroups), or on respondent behavior. Nothing concerning non-respondents is captured from this indicator. Thus, in web surveys research into non-sampling error in a variety of aspects is needed.

The key point is that the technology used on the Web as a survey tool inherently contains certain advantages for monitoring and studying survey errors. It allows for the collection of so-called “para-data”, i.e., data that accurately describes the interviewing process and allows for the linkage of the collected data with certain characteristics of the interviewed people. Thus, in addition to the statistical approaches related to bias and estimation, it is possible to develop behavioral models, audit trail models and various kinds of models and indicators which could help in defining the best strategies at different points in the survey process.

2.1 *Bias due to under-coverage and non-response*

In this paragraph, I first give a brief introduction to the concepts of under-coverage and non-response related to bias. Following this, bias related to under-coverage and non-response is formally defined and some estimation criteria are presented. At the end of the paragraph some experimental studies and some possible new experimental approaches related to modeling non-response and estimating bias are proposed.

We face a coverage error when the sample frame is not appropriate to the population for which the survey research was intended. This error is expressed as members of the survey population not having a known non-zero (often equal) chance of being included in the sample (Dillman, 2007). Normally, what is experienced is that some appropriate members are not included in the sample frame and thus have a zero chance of being selected for the survey. Under-coverage might occur in a web survey if the target population is wider than just persons with Internet access and respondents are selected via the Internet. Therefore it is impossible for people without the Internet to participate in the survey. This type of under-coverage can have serious consequences. If people without Internet access systematically differ from persons with access to the Internet, estimates of population parameters may be seriously biased.

Non-response error is a function of two components, both of the rate of non-response and of the differences between respondents and non-respondents on the variables of interest (Groves and Couper, 1998). This last component is especially critical in Web surveys. The non-response problem concerns missing data for some units of a sample, either partially, or for all variables. Literature ([Groves and Couper, 1998](#)) usually differentiates between unit and item non-response. Unit non-response refers to the complete loss of a survey unit (due to inaccessibility, refusal, or inability to respond), while item non-response (when surveys are partially completed and

returned) refers to missing responses to individual questions. An additional category is complete response. This paper is concerned with the issue of unit non-response.

Basically, non-response would not present a problem if it were randomly distributed. However, to the degree that interest lies with issues underlying self-selection, we can expect that variables affecting the decision to participate could also affect the answers to the variables in the questionnaire. Consequently, the bias of self-selection will then lead to results that are not representative for the population as a whole, thus creating a non-response bias. Different non-response components could affect the bias in different ways (the unknown characteristics and attitudes of non-respondents may cause inaccuracies in the results). The primary goal when investigating non-response is to validate the obtained responses by demonstrating that the non-response is either insignificant in size or random (and thus not causing bias), although in most cases information is not randomly missing. Let us now consider weighting techniques that could help in reducing bias in Web surveys.

Let us assume that it is the task of the Web survey to estimate the mean $\bar{y}_I$ of the Internet population, Bethlehem (2002) shows that the bias of the sample mean $\bar{y}_S$ can be defined as

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y}_I = \frac{R_{pY} S_p S_Y}{\bar{p}} \quad (1)$$

where R_{pY} is the correlation coefficient between the target variable and the response behavior, S_p is the standard deviation of the response probabilities and is the S_Y standard deviation of the target variable.

In Web surveys some weighting adjustment techniques should be applied for non-response and under-coverage. In Biffignandi and Bethlehem (forthc.) it is shown that post-stratification and calibration as theoretical framework for generalized regression estimation and raking ratio estimation might reduce bias. As is well known, weighting adjustment requires auxiliary variables. The existence of a strong relationship between survey and auxiliary variables is central to success in reducing bias.

If the auxiliary variable population distribution is not available, we might use a reference survey to collect this information.

Bethlehem (2007) has shown that the variance of the post-stratification estimator in the presence of reference survey data is:

$$V(\bar{y}_{I,RS}) = \frac{1}{m} \sum_{h=1}^L W_h (\bar{Y}_I^{(h)} - \bar{Y}_I)^2 + \frac{1}{m} \sum_{h=1}^L W_h (1 - W_h) V(\bar{y}_I^{(h)}) + \sum_{h=1}^L W_h^2 V(\bar{y}_I^{(h)}) \quad (2)$$

If the quantity $\bar{y}_I^{(h)}$ is measured in the Web survey, their variance $V(\bar{y}_I^{(h)})$ will depend on n (the Web survey size), which in practice will generally be much larger than m (reference survey size, which is usually small). To sum up, the first term in the variance will be the most important one, i.e., the large number of observations in the Web survey does not help to produce accurate estimates. Using a reference survey may well reduce the bias of estimates, but at the cost of a higher variance.

If the distribution of the auxiliary variables in the population is available and auxiliary variables are correlated with the target variable post-stratification, generalized regression and ranking ratio estimation can be effective in reducing variables, as demonstrated in Biffignandi and Bethlehem (forthc.). As regards the estimation techniques proposed by Deville and Särndal (1992) and Deville et al. (1993) if there is non-response, under-coverage or self-selection, the asymptotic properties are not equal. In other words, whether the bias can be removed or reduced will depend on how well the chosen underlying model is estimated on the basis of the observed data.

Further research necessitates empirical studies based on alternative methods. For instance, an interesting approach, worth experimenting with, is the extension for non-response estimation methods with follow up surveys, as proposed by Chambers and Seppo (2006). A current strategy for dealing with non-response problems is to reweight the respondents' data to account for their different response propensities. However, as discussed above, this depends on the availability of auxiliary information that "explains" the non-response, and implicitly assumes that the non-respondents' data are missing at random (MAR) given this auxiliary information. Clearly, the MAR assumption cannot be tested simply using the respondents' data. In situations where the MAR assumption seems unjustified, therefore, one option is to carry out a follow-up of a subsample of non-respondents. This survey aims at building a "better" overall model for response propensity (e.g., by incorporating information obtained from the sampled non-respondents) as well as allowing a more sophisticated method of estimation that uses the data obtained from both respondents and sampled non-respondents. Independently from comparing to formalization of this approach to the others, empirical analysis is important for suggesting best survey strategies and highlighting the impact of respondents and non-respondents on the estimation results. One survey for which the follow-up approach is interesting is the one on innovation since innovative and non-innovative businesses could greatly differ in their characteristics as well as in their response propensity.

2.2 *Para-data and survey participation*

One of the substantial advantages of Web surveys compared to mail surveys is that information is automatically collected during the survey process in addition to responses to the substantive questions. These data are called 'para' or 'meta-data'.

More precisely, two kinds of information are collected, by tracing automatically,:

- 1) the response process including the use of cgi scripts, java applets, and log files; includes characteristics of a respondent's browser captured from the server logs (server-side paradata);
- 2) the answering process within the questionnaire, i.e., insight into the sequencing and completeness of responses. Respondent behaviour is traced on each Web page as they answer the survey (client-side paradata) by embedding JavaScript code in survey instruments, including such measures as response latencies (e.g., time to response, click), order of responding, changing of answers, mouse movements, and the like.

Due the huge amount of available information it is no longer necessary to view responses to survey requests within the confines of three generic behaviors (unit non-response, item non-response, complete response). Such information support the

existence of many possible responses to requests for survey participation. Typologies of response behavior to explain more fully the potential variations in participation in Web-based surveys have been introduced in the literature and can provide a valuable insight into understanding non-response and various response behavior. Regardless of the specific approach used, the data allow the reconstruction of the response patterns in Web surveys. Several steps of the Web survey process can be analysed and the final response rate is divided into different components, one for each step. For example, for telephone pre-recruited Web surveys of Internet users, the following steps can be identified following telephone pre-recruitment (telephone interview, e-mail address collection); e-mail recruitment (sending out e-mail invitations, absorption (acceptance) of e-mails by the network), access to the questionnaire, and the questionnaire completion step. A further approach for investigating participation in web surveys is based on the analysis of the timeliness of the participant's survey response and on the features of participants' Internet usage. The approach permits measurement of the times of reaction to e-mail invitations and the modeling of survival in the survey process. Survival in the survey process depends mainly on the features of participants' Internet usage: purpose of Internet usage and place of connection seem to be the key factors in the speed of reaction of eligible respondents.

An overview of the stages in the participation process for Web surveys in general, both for individuals and establishments, is given by Biffignandi and Pratesi (2002), and by Pratesi, Lozar Manfreda, Biffignandi, Vehovar (2004).

A conceptual model of the Web survey process based on the following stages has been proposed by Fan and Yan (2010). It identifies the following stages: survey delivery, survey completion and survey return. The model is used to systematically review a wide variety of factors influencing the response rate in the stages of survey development.

It should be underlined that initially researchers collected para-data to study the response propensity of different groups and, consequently therefore to improve the design or performance of the instruments, to understand, for instance, where resources and efforts should be directed to improve response rates. Increasingly they are now being used to understand or even control respondent behavior. An interesting new area of research now concerns the application of methods and concepts from cognitive psychology to the development of new computerized methods of data collection, and the increase in concern with regard to measurement and non-response as sources of error in survey estimates. In the follow I refer to this research areas as behavioral analyses.

2.3 Behavioral analyses

This area of research is based on a social scientific mode and fits well into the new paradigm, which focuses on non-sampling error, besides sampling error. In this approach, the focus of research turns toward understanding why some respond to Web surveys while others do not, and whether and in what ways these two groups may differ on key variables of interest. The recent development of client-side para-data allows us to better understand the way respondents construct their answers by providing new types of data, including time taken to answer a question and whether respondents change their answers. The results suggest that the visual layout of survey questions not only affects the number but also the types of changes respondents make. The theory

does not concern itself with the impact of non-response on estimates although they have implications for the question of when non-response is likely to bias result.

Theories aim to answer questions such as: who is more likely to respond, why non-response occurs, who is likely to be hard to reach, and how interest in the survey topic affects the willingness to participate.

From the methodological point of view behavioral analyses are related to the Cognitive Aspects of Survey Methodology movement (CASM) and in many empirical studies the TPB (theory of planned behavior, Ajzen, 1991) model is applied. For instance, Tourangeau (2003) proposes an extended TPB model to explain the intentions of potential respondents in participating in Web-based surveys. The chief objective is to obtain a more comprehensible picture of how intentions are formed. Hierarchical regression and structural equation modeling is used and the results indicate that attitude of potential respondents, social norm, perceived behavioral control and PWI are determinants of unit non-response in Web surveys. The results highlight the usefulness of the TPB model in explaining the intention of potential respondents in taking part in Web surveys; also the importance of the role of *PWI* in Web surveys is verified.

Responding to suggestions of prior research in examining the psychological constructs involved in a decision for or against participation in Web surveys, the effects of trust in sponsor and personal innovativeness on potential respondents' intention to participate have been also investigated. Based on the theory of planned behavior (TPB), two alternative models were empirically tested, in which the roles of trust and innovativeness were theorized differently - either as moderators of the effects that perceived behavioral control and attitude have on participation intention (moderator model) or as direct determinants of the attitude, perceived behavioral control and intention (direct effects model).

In its simplest form, the Theory of planned behavior can be expressed as the following function:

$$BI = (W1)AB[(b) + (e)] + (W2)SN[(n) + (m)] + (W3)PBC[(c) + (p)] \quad (3)$$

BI : Behavioral intention, AB : Attitude toward behavior, (b): the strength of each belief, (e): the evaluation of the outcome or attribute, SN : social norm; (n): the strength of each normative belief, (m): the motivation to comply with the referent, PBC : Perceived Behavioral Control, (c): the strength of each control belief, (p): the perceived power of the control factor, W : empirically derived..

2.4 Concluding remarks

Given the above-mentioned issues, non-sampling errors represent quite an interesting and wide area for new research, especially in the context of Web surveys.

Although the idea of developing synthetic indicators for total error is attractive, Web survey methodology, and the consequent quality of the statistical economic data collected, could be considerably advantaged and improved even by studying some specific aspects of non sampling errors. Generally speaking, methodology and quality of the socio-economic collected data could greatly benefit by:

- a) the development of suitable estimation methods aimed at capturing the bias and specific variance connected with the frame characteristics and participation process of this type of survey;
- b) research, principally based on experimental designs, allowing the effects of various factors to be tested (for example, the effects of different types of question structure, various contact modes, etc.);
- c) research, based on behavioral models, that allows response and participation processes to be analyzed in the context of the individual behavior of survey respondents.

The recent past has already seen progress in the analysis of non-response, some of which – focusing on Web surveys - has been described briefly in this paper. However, there still remain many open questions, touching on a whole variety of aspects.

References

1. Ajzen L. (1991): The theory of planned behavior. *Org. Behav. Human.* R0179. 211
2. Bethlehem, J.G. (2007): Reducing the Bias of Web Survey Based Estimates. Discussion Paper 07001. Statistics Netherlands, Voorburg/Heerlen
3. Bethlehem, J.G. (2002): Weighting nonresponse adjustments based on auxiliary information. In: Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (Eds): *Survey Nonresponse*. Wiley, New York, USA
4. Biemer P., Lyberg. L. (2003): *Introduction to Survey Quality*, Wiley
5. Biffignandi S., Bethlehem J.: *Handbook for Web Surveys*, Wiley, forthcoming.
6. Biffignandi S., Pratesi M.(2002): Internet surveys: the role of time in Italian firms response behaviour, *Research in Official Statistics*, n. 2
7. Biffignandi S., Toninelli., Fabrizi E. (2010): The Effects of Non-Response Reweighting on Estimators' Precision in a Web Survey, 45th SIS Scientific Meeting
8. Chambers R., Seppo L.: Survey Estimation Under Informative Non response with Follow-up, *Journal of Official Statistics*, Vol. 22., N.-1 (2006)
9. Couper M. (2001): Web survey research: challenges and opportunities M. P. Couper, *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9
10. Deville, J.C., Särndal, C.E. (1992): Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association* 87, pp. 376-382
11. Deville, J. C., Särndal, C. E. & Sautory, O. (1993): Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88, pp. 1013-1020.
12. Dillman D. (2007): *Internet Mail and Internet surveys, The Tailored Design Method*, Wiley
13. Fang J., Shao P., Wan J. (2007): Intention to Participate in Web Surveys: Extended TPB Model, *Proceedings of the WiCom 2007 Conference*, Shangai
14. Groves R., Couper M. (1998): *Nonresponse in Household Interview Surveys*, Wiley
15. Platek R., Särndal C.E.: Can a Statistician deliver?, *Journal of Official Statistics*, vol. 17, n.1 (2001)
16. Pratesi M. , Lozar Manfreda K., Biffignandi S., Vehovar V. : List-based Web Surveys: Quality, Timeliness, and Nonresponse in the Steps of the Participation Flow, vol. 20, n. 3 (2004)
17. Tourangeau R. (2003) Cognitive Aspects of Survey Measurement and Mismeasurement, *International Journal of Public Opinion Research* 15:3-7