STEPHANIE STEINMETZ
KEA TIJDENS

Can weighting improve the representativeness of volunteer online panels? Insights from the German Wage indicator data

Introduction

During the last decades, the web has become a popular tool of data collection not only for commercial marketing agencies but also for scientific purposes. The introduction of web surveys has triggered a heated debate about their scientific validity (Fricker and Schonlau 2002, Ilieva et al. 2002). Arguments in their favour emphasise cost benefits, fast data collection, ease of processing results, flexibility of questionnaire design, and the potential to reach respondents across national borders. Arguments against web surveys mainly focus on traditional types of survey errors and related questions of quality, reliability and validity for scientific use.

These concerns relate to various sources of survey error. At present, the *coverage error* is a serious problem for many web surveys, particularly those targeting the general population. As not every person has Internet access, and a list of e-mail addresses covering the whole population does not exist, the probability of being included in the survey is not the same for everybody. Moreover, though Internet penetration rates continue to increase, the possible bias is related not only to the number of people having access to the Internet, but also to the differences among them in age, gender, education, and behavioural characteristics (Dever et al. 2008). Another major difficulty concerns implementing a probability-based web survey in the absence of an adequate sampling frame (Couper 2000). Problems arise particularly when adopting *non-probability* and *self-selection* recruitment methods. In this case respondents form a convenience rather than a probability sample and little is known about the degree to which the obtained results can be generalised for the whole population. Furthermore, people who self-select into a survey may differ from those who do not in terms of time availability, web skills, or altruism to contribute to the project (Fricker 2008).

This is particularly problematic for non-probability based web surveys, like volunteer online panels.[1]. Once a sample of potential respondents has been selected, the methodological concerns continue. Not all sample members may be willing or able to complete the survey. The extent of bias depends on the rate of *non-response* as well as on differences between respondents and non-respondents with regard to the variables of interest. Non-response bias is not unique to web surveys. But as their response rates tend to be lower when compared to other modes (Lozar Manfreda et al. 2008), the problem is particularly severe. Moreover, for non-probability web surveys, the problem of non-response is hard to define because its evaluation is traceable only in cases where the frame and the chance of selection are known.

II. Different weighting techniques as possible solutions

As indicated above, a representative sample of the population is of paramount importance when conducting a survey. The under- or overrepresentation of certain socio-demographic characteristics within the collected sample introduces bias and affects the reliability of the results. Comparing the population distribution of a variable with its sample distribution, it can be assessed whether the sample represents the population adequately with respect to this variable. If the distributions vary considerably, the sample is selective. To correct this, adjustment weights can be computed to align, in the sample, the distribution of the selective variable with the distribution observed in the population. There are several methods allowing this adjustment. In the present context, *post-stratification weighting* and *propensity score adjustment (PSA)* will be described in more detail.

1 Coverage and sampling is less of a problem where all members of the target population use the Internet and where the e-mail addresses of all members are known, like in the case of students or members of organizations. Here, the existence of a proper sampling frame allows the drawing of a probability-based sample and the generalising of conclusions to the whole population using standard inference procedures.

Post-stratification weighting is one of the common methods. It has mainly been applied to correct socio-demographic differences between the (web) sample and the population under consideration. The formula for such weights w_i is:

(1) $w_i = p_p / p_s$

Where p_p is the population proportion, and

p_s is the sample proportion.

Propensity Score Adjustment (PSA) was originally developed for the comparison of populations in the context of experimental designs (Rosenbaum and Rubin 1984). It aims to correct differences in socio-demographic and 'webographic' (attitudinal or behavioural) variables regarding individual decisions to participate in web surveys (Lee and Vaillant 2009, Loosvedt and Snock 2008, Schonlau et al. 2009). A crucial step in designing PSA is the identification of potentially relevant covariates which capture the difference between the web respondents and the population of interest.[2] To apply PSA, a probability-based reference survey is needed which contains the required covariates. After merging both samples using the variables common to both data sets, an indicator variable (I_i) is defined indicating whether the respondent belongs to the web survey or not. Subsequently, the web sample is adjusted to the probability sample by estimating via a logistic regression the probability of each respondent to participate in the web survey using a set of selected covariates (X_i) as predictors.

$$(2) \quad \log\left(\frac{\tau(X_k)}{1-\tau(X_k)}\right) = \alpha_k + \beta_k^{'} X_k + \varepsilon_k$$

After calculating the propensity scores, the next step is to balance the non-equivalent groups using matching, stratification, covariance adjustment, or weighting on the estimated propensity score. When applying *propensity score weighting,* weights (w_i^{ps}) are formed as the inverse of the propensity score (Schonlau et al. 2007). Since the propensity scores refer to both web and reference survey respondents, the propensity score weights for the two samples (I_i=0 and I_i=1) are as follows:

$$(3) \quad w_i^{ps} \begin{cases} 1/\,ps_i & \text{if } I_i = 1 \text{ (web survey)} \\ 1/(1-ps_i) & \text{if } I_i = 0 \text{ (reference survey)} \end{cases}$$

2 Applications of PSA for web surveys show that all researchers adjust for socio-demographic variables. Moreover a variety of specific variables, such as self-assessed health status (Schonlau et al. 2007), and self-rated social class, employment status, political party affiliation, having a religion and opinion towards ethnic minorities (Lee 2006), has been used for propensity score adjustments.

Against this background, it becomes evident that the development of sample weights plays an important role in the correction of selection bias in (non) probability-based web surveys. It has been emphasised that for generalising web survey results for the whole population, post-stratification and propensity-based weights are necessary (Duffy et al. 2005). Nevertheless, the implications of the different adjustment procedures are still under discussion. Until now their application has produced rather diverse results, and there is no certainty as to whether the representativeness of web surveys can be improved (Taylor 2005, Vehovar et al. 1999). Therefore, the next section will evaluate the effectiveness of the two described weighting techniques in adjusting biases arising from non-randomised sample selection using the 2006 German *WageIndicator Survey* data (Lohnspiegel)[3] and the Socio-Economic Panel (SOEP) as a probability-based reference survey. The SOEP is used as a reference survey because it provides the greatest overlap of variables with the Lohnspiegel.

III. Analyses

Both data sets collect information on a wide range of subjects including basic demographics, wages and other work-related topics. Most importantly, they also include variables such as health and job satisfaction, which can be considered as webographic variables. Furthermore, as the applied weights should improve all kind of wage estimations, the samples have been restricted to employees, persons aged between 16 and 75 living in Germany, and a monthly gross wage between 400-10000€. For PSA it was also necessary to eliminate missing values which finally led to samples of N=21914 (Lohnspiegel) and N=7993 (SOEP).

As indicated above, before weighting techniques can be implemented, it is important to evaluate the selection bias by comparing the distributions of specific variables between the web and the reference survey. In this respect, research has shown that the Lohnspiegel is affected by typical selection bias: women, highly educated persons[4], older persons (45-65+), part-timers, persons in manual occupations, persons living in a region with a high unemployment rate, and persons who are satisfied with their health and their job are underrepresented (Steinmetz and Tijdens 2009).

3 Although in most countries the number of observations of the WIS is larger than in national labour force surveys, the samples seem to fail to be representative of the population because of the above-mentioned methodological problems.

4 This is a rather surprising finding that might be due to the fact that in Germany persons with 'lower' education might be attracted to the project via a web-link which is prominently placed on a trade union-related homepage.

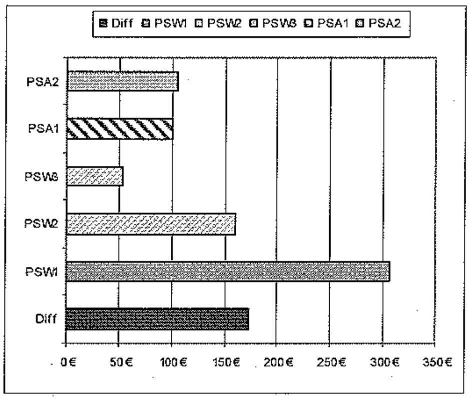
Based on these findings, post-stratification and PSA weights are constructed. The first post-stratification weight (PSW1) simply contains the core demographics variables (gender, education and age). However, as the description revealed that part-timers are also underrepresented, the second weight (PSW2) includes gender, education, age and part time. Finally, the third weight (PSW3) examines whether a minimum of variables (only part time and job satisfaction), which are predominantly affected by selection bias, is sufficient to adjust the samples.

For PSA, different models have been defined. The first model (PSA1) includes gender, education, age, occupation, working hours, type of contract, region, and log wage. To test the effect of 'webographic' variables, the second model (PSA2) additionally includes variables on health and job satisfaction. Running logistic regressions, for all model specifications the variables are found to have a significant effect on the selective participation in the web survey (see for more detailed information on the analyses, Steinmetz and Tijdens 2009).

Results

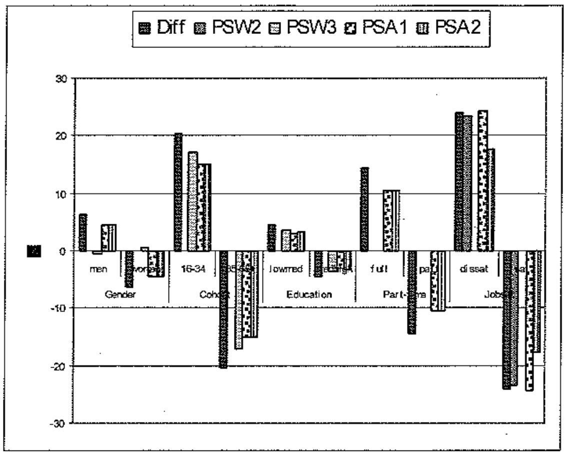
The following figure 1 describes the differences in the mean wages between the unweighted and weighted *Lohnspiegel* and the SOEP. The black bar (Diff) indicates the differences between the two data sets without weighting. It shows that the mean wage in the *Lohnspiegel* is around 173€ higher than in the SOEP.

Figure 1: Differences in gross monthly mean wages, weighted and unweighted LS and SOEP, 2006



Source: German LS and SOEP 2006, own calculations

Figure 2: Differences in the effectiveness of weights for selected variables, weighted and unweighted LS and SOEP, 2006



Source: German LS and SOEP 2006, own calculations

A properly designed post-stratification and propensity score weight could be expected to reduce the difference between the two data sets.

The application of the different weights indicates that out of the three post-stratification weights, only PSW2 (gender, education, age and part time) and PSW3 (part time and jobs satisfaction) are capable of adjusting the mean income of the web sample to the reference sample (striped bars). Also the two defined propensity score weights (PS1-2) reduce the differences between the web and the reference survey. However, PSA1 (containing only socio-demographic and labour market-related variables) seems to be more efficient in adjusting the two samples.

In a next step, the four 'successful' weights were implemented for the adjustment of the distribution of selected variables. Figure 2 shows the differences in the percentage of these variables between the SOEP and the unweighted (black bar) and weighted *Lohnspiegel*. Again, it is expected that the differences between the two unweighted surveys weaken with the implementation of the different weights. The results indicate that the application of post-stratification weights is capable of levelling out the under- and overrepresentation of the combined classes of gender, age, education and part-time (PSW2) or part time and job satisfaction (PSW3). However, it does not enhance the comparability of web survey and reference survey respondents, for instance, with regard to their job satisfaction (when applying PSW2). With respect to the propensity score weights, the results show that the differences between the SOEP and the *Lohnspiegel* became slightly smaller for all variables.

Only in case of job satisfaction the results for PSA1 (without any webographic variable) indicates that the differences even became larger. However, in comparison to the post-stratification weights, propensity score weights do not lead to a complete adjustment of the two samples with respect to specific variables.

IV. Conclusion and Discussion

Besides several arguments in favour of web surveys, there are, particularly for non-probability based web surveys, fundamental concerns about the generalisation of results from such data sources for the whole population. In this contribution, post-stratification and propensity score weights have been applied to a German web sample in order to determine the extent to which they are capable to correct observed selection biases. Similar to findings from previous studies (Lee 2006, Schonlau et al. 2009), the results for post-stratification weights show that, on average, the impact is very limited. Furthermore, they *'make the proportions of the variables used comparable, but this does not necessarily make the answers between web respondents and personally interviewed people more comparable with regard to attitude questions'* (Loosveld and Sonck 2008, p. 104). Only minimal changes could be observed when applying PSA to adjust also for attitudinal questions. It should moreover be underlined critically that the difference between the samples with respect to selected variables of interest sometimes even became larger instead of smaller (see PSA1). Moreover, the inclusion of additional 'webographic' variables did not improve the adjustments considerably.

Against this background, the findings illustrate that the different weighting methods using balancing variables do not fundamentally increase the representativeness of web survey data. Even though the applied weighting techniques do not seem to provide adequate possibilities of generalising web survey results for the whole population, it should be pointed out that the collected 'unweighted' web data are not useless. The underlying reasons for the failure of the applied weights could have different reasons: the used reference survey might also be affected by selection bias, there may be different mode effects in the web and the reference sample (web questionnaire vs. face to face interviews). With regard to propensity scores, it could also be argued that more variables, particularly webographic variables, have to be included. To clarify these problems, more analyses and advanced correction techniques are needed.

Finally, critics stressing problems of generalising findings of non-probability based web surveys should consider that the problem of self-selection may also arise in the case of probability-based samples. Persons who are willing to participate in a survey always differ from those who are not participating. In this context, it seems worthwhile to recall the advice by Couper and Miller (2008) not to treat survey quality as an absolute goal, but to evaluate it relative to other features of the design and the stated goals of the survey.

References

Couper, M. (2000): Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, Vol. 64 (4), pp. 464-481.

Couper, M.; Miller, P. (2008): Web survey methods - Introduction. *Public Opinion Quarterly*, Vol. 72(5), pp. 831-835.

Dever, J.; Rafferty, A.; Valliant, R. (2008): Internet surveys: can statistical adjustments eliminate coverage bias? *Survey Research Methods,* Vol. 2(2), pp. 47-62.

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

Fricker, R. (2008): Sampling methods for web and Email surveys, pp. 195-216. N. Fielding; R. Lee; G. Blank (eds.), *The SAGE Handbook of Online Research Methods*. London: Sage.

Fricker, R.; Schonlau, M. (2002): Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods*, Vol. 14(4), pp. 347-367.

Ilieva, J.; Baron, S.; Healey, N. (2002): Online surveys in marketing research: pros and cons. *International Journal of Marketing Research,* Vol. 44(3), pp. 361-76.

Lee, S.; Valliant, R. (2009): Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods Research,* Vol. 37(3), pp. 319-343.

Lee, S. (2006): Propensity Score Adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, Vol. 22(2), pp. 329-349.

Loosveldt, G.; Sonck, N. (2008): An evaluation of the weighting procedures for online access panel surveys. In: *Survey Research Methods*, Vol. 2(2), pp. 93-105.

Lozar Manfreda, K.; Bosnjak, M.; Berzelak, J.; Haas, I.; Vehovar, V. (2008): Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, Vol. 50(1), pp. 79-104.

Rosenbaum, P.; Rubin, D. (1983): The central role of the propensity score in observational studies for causal effects. *Biometrika*, Vol. 70(1), pp. 41-55.

Rosenbaum, P.; Rubin, D. (1984): Reducing bias in observational studies using sub classification on the propensity score. *Journal of the American Statistical Association,* Vol. 79(387), pp. 516-524.

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(3), pp. 291-318.

Schonlau, M.; van Soest, A.; Kapteyn, A. (2007): Are ‚Webographic' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Metods,* Vol. 1(3), pp. 155-163.

Steinmetz, S; Tijdens, K. (2009 forthcoming): Comparing different weighting procedures for volunteer web surveys: Lessons to be learned from German and Dutch Wage indicator data. *AIAS Working Paper*, University of Amsterdam.

Taylor, H. (2005): Does Internet research 'work'? Comparing online survey results with telephone surveys. *International Journal of Market Research*, Vol. 42(1), pp. 51-63.

Vehovar, V.; Manfreda, K. (1999): Web surveys: can the weighting solve the problem? *Proceedings of the Survey Research Methods Section, American Statistical Association (1999)*, pp. 962-67. See: http://www.amstat.org/sections/srms/Proceedings/y1999f.html

Stephanie Steinmetz works as a Post-doc researcher in the project "Improving web survey methodology for social and cultural research" at the Erasmus Studio Rotterdam. Her main tasks include the study of methodological issues concerning web surveys, particularly selection bias, item-non response and weighting procedures. Email: steinmetz@fsw.eur.nl

Professor Kea Tijdens is a Research Coordinator at Amsterdam Institute of Advanced Labour Studies (AIAS) at the University of Amsterdam, and a Professor of Women and Work at the Department of Sociology, Erasmus University Rotterdam. She is scientific coordinator of the continuous WageIndicator web-survey about work and wages, which is currently operational in 45 countries. She has coordinated EUROCCUPATIONS, an EU-FP6 project aiming to build an 8-country occupations database (2006-2009). Current research interests include measurement of occupations, industrial relations, collective bargaining and the upgrading and downgrading of occupations. She has published in Feminist Economics, Journal of Family Issues, Information & Management, Acta Sociologica, New Technology, Work and Employment, and Economic and Industrial Democracy. Email: K.G.Tijdens@uva.nl