MATHEMATICAL POPULATION STUDIES
2017, VOL. 24, NO. 3, 161–171
https://doi.org/10.1080/08898480.2017.1330012

Routledge
Taylor & Francis Group

Using experts' consensus (the Delphi method) to evaluate weighting techniques in web surveys not based on probability schemes

Vera Toepoel and Hannah Emerson

Methods and Statistics, Utrecht University, The Netherlands

ABSTRACT

Weighting techniques in web surveys based on no probability schemes are devised to correct biases due to self-selection, undercoverage, and nonresponse. In an interactive panel, 38 survey experts addressed weighting techniques and auxiliary variables in web surveys. Most of them corrected all biases jointly and applied calibration and propensity score adjustments. Although they claimed that sociodemographic and web-related variables are the most useful auxiliary variables to employ in adjustments, they considered only sociodemographic variables to correct biases because of their availability.

KEYWORDS

auxiliary variables; Delphi study; nonprobability samples; web surveys; weighting

1. Introduction

Web surveys are easy to create, inexpensive, time efficient, and can reach millions of Internet users. However, they often fail to be representative of the population because they have no sampling frame and lack a probability-based sampling procedure necessary to correct for biased estimates. Bethlehem and Biffignandi (2012) and Baker et al. (2013) state that biases may occur in web surveys due to self-selection, undercoverage, nonresponse, and sampling errors. Weighting is used to adjust survey statistics for unequal selection probabilities, coverage error, or to represent population characteristics on several covariates. Lee (2006), Pedrazza et al. (2007), Malhorta and Krosnick (2007), and Steinmetz et al. (2014) question the reliability of adjustment techniques in web surveys, be they designed to correct self-selection, undercoverage, and nonresponse together or separately. Similarly, the choice of auxiliary variables in web surveys has been addressed theoretically but insufficiently in the objective of constructing unbiased estimates.

We use the Delphi method, which relies on experts' consensus (Powel, 2003). A three-wave survey was sent to experts belonging to different academia, government, and market research agencies. These experts were to reach

CONTACT Vera Toepoel v.toepoel@uu.nl Utrecht University, Padualaan 14 3584CH Utrecht, the Netherlands.
Color versions of one or more of the figures in the article can be found online at ww.tandfonline.com/gmps.

a consensus about the appropriate weighting technique to be used with web surveys. We raised three questions:

(1) Do experts correct for biases separately or altogether?
(2) What weighting techniques do experts use in web surveys?
(3) What auxiliary variables do experts use in adjustments and which variables do they find the most efficient?

2. Background

Biases in web surveys typically result from self-selection, undercoverage, nonresponse, or sampling error. Self-selection error occurs when participants have volunteered to participate, which can lead the sample to misrepresent the population. Undercoverage occurs in web surveys when not all individuals of the sample have been reached. This error is common in web surveys because not everyone has an email address. Nonresponse can be due to unwillingness or inability to participate. Nonresponse error occurs when nonrespondents differ from the population with respect to variables of interest.

Bethlehem and Biffignandi (2012) have shown that weighting is often necessary for samples not based on probability schemes. In samples based on a probability scheme, adjustments are based on the inverse of the probabilities of selection, which is not possible in samples that are not based on probability schemes. Weighting techniques used in surveys based on no probability schemes rely on auxiliary variables, whose values are known for the population. These variables take values proportional to those taken by the population variables. For example, for a sample of 100 people (30 men and 70 women) and for a population comprising 48% men and 52% women, men's responses are weighted more than women's responses to make the sample proportional to the population with respect to sex. In this example, the weight for men equals 48/30=1.6 and the weight for women equals 52/70=0.74. By using these weights, one hopes to reduce the bias between the estimate produced by the sample and the population estimate.

There are several techniques for applying these weights. The two most known techniques are poststratification and adjustments by propensity score. Other commonly used techniques include estimation by raking ratio and regression. Matching is a pre-survey sampling method, and reference survey estimation is usually used only as a means of calculating propensity scores in the frame of an adjustment by propensity score.

Poststratification, also known as cell weighting, requests to distribute the population into strata or cells using auxiliary variables, such as age, sex, or education level, in order to create a representative sample with respect to the auxiliary variables (Kalton and Flores-Cervantes, 2003). People in a given

stratum are weighted by the ratio of the population percentage to the sample percentage in that stratum. The sample should be proportional to the population in each stratum by conforming the sample joint distribution of the auxiliary variables to that of the population. Kalton and Flores-Cervantes (2003) argue that poststratification weighting is often used when little is known about nonrespondents and auxiliary information is limited. However, poststratification weighting can produce poor adjustments when strata are few or samples are small. Alternative methods perform better when auxiliary information is available.

Deming and Stephan (1940) devised estimation by raking ratio, which is a variant of poststratification (Lee, 2006). Poststratification relies on the joint distributions of the auxiliary variables; raking relies on their marginal distributions. Through iterative proportional fitting (Deming and Stephan, 1940), the distribution of the population resulting from estimation by raking ratio uses the marginal distributions of the auxiliary variables and correlations. Weights are then calculated from the joint population distribution. The estimates have smaller variances by raking than by poststratification, but the weights may also differ between the two methods. Raking can be used when only marginal information is known per stratum or weighting is possible only with joint distributions. For few strata and large sample sizes, poststratification weighting may be better. However, for numerous strata (from many auxiliary variables) and a small sample, Kalton and Flores-Cervantes (2003) advocate raking.

Regression estimation involves neither joint nor marginal distributions but confirms sample estimates to population parameters. Regression is based on a linear regression model first introduced by Deville and Sarndal (1992). The sample estimate is equated to the population total output, and the weights are devised to fit for the population totals. They can be seen as regression coefficients (Kalton and Flores-Cervantes, 2003). As with standard regression, variables can be modified and interactions specified.

Adjustment by propensity score is based on logistic regression (Rosenbaum and Rubin, 1983). Logistic regression was used in weighting adjustments for nonresponse (Lee, 2006). Propensity scores are the probabilities of participating in the survey. Usually a logistic regression model is used in which the indicator variable is the dependent variable and attitudinal variables are the explanatory variables. Weights are inverse propensity scores. Kalton and Flores-Cervantes (2003) equate adjustment by propensity score to raking when auxiliary variables are categorical with no interactions. They claim that it is more flexible than raking for its ability to include continuous predictors. It can also cope with interactions, using joint (as in post-stratification) and marginal (as in raking) information.

The quality of the adjustment depends on the auxiliary variables. Pedrazza et al. (2007) found that poststratification reduces biases only partially. Dever et al. (2008) recommend using age, ethnicity, sex, education, presence of

children in household, employment, and marital status as auxiliary variables. Schonlau et al. (2007) found that web-related attitudinal and behavioral questions complete those about demographics efficiently to produce better representative samples. We found that weighting adjustments fail to erase all differences between online and offline survey results (Duffy et al., 2005; Taylor, 2005; Malhorta and Krosnick, 2007; Pedrazza et al., 2007; Loosveldt and Sonck, 2008; Scherpenzeel and Bethlehem, 2010).

3. Method

3.1. *The Delphi method*

The Delphi method is devised to obtain consensus among experts (Delbecq et al., 1975; Powel; 2003) through an iterative procedure over several questionnaires. The responses of experts to a first questionnaire are combined. Another questionnaire is sent out with feedback from the first questionnaire and so on until a consensus among experts is obtained. The participants in a Delphi study are usually selected through administration lists or for their reputations. Okoli and Pawlowski (2004) advise to take between 10 and 18 experts. The experts need not meet face-to-face, and they can remain anonymous so they can be more comfortable speaking on sensitive issues.

3.2. *Participants*

We contacted at least 10 experts in academia, government, and market research from the American Association for Public Opinion Research, the European Survey Research Association, and the International Sociological Association Research Committee on Logic and Methodology.

The first wave consists of 37 experts. Delphi studies are prone to attrition, whereby experts do not complete all waves, so we recruited additional experts for the second wave, but stuck to the same experts for the third wave. The second wave has 38 experts and the third wave 36. Twenty-one experts completed all three waves. At the third and last wave, 25 experts were from Europe and 11 from North America, with a mean age of 42.1 years (standard deviation =11.1), with 27 men and 9 women. Two experts had 0–5 years of research seniority; 12, 5–10 years; 9, 10–15 years; and 13, over 15 years. Eight experts were from government, 13 from market research, and 15 from academia. Table 1 presents the characteristics of the experts at each wave.

3.3. *Questionnaires*

At the first wave, experts could express their opinions on weighting techniques in open-ended questions. The first wave consisted of 52 questions

Table 1. Characteristics of experts.

Variables	Wave 1	Wave 2	Wave 3
Average age (in years)	42 (sd=13)	43 (sd=12)	42 (sd=12)
Sex			
Man	28 (76%)	28 (74%)	27 (25%)
Woman	9 (24%)	10 (26%)	9 (25%)
Sector			
Academia	25 (68%)	16 (42%)	15 (42%)
Government	4 (11%)	9 (24%)	8 (22%)
Market research	8 (21%)	13 (34%)	13 (36%)
Seniority			
0-5 years	5 (14%)	2 (5%)	2 (6%)
5-10 years	9 (24%)	13 (34%)	12 (33%)
10-15 years	9 (24%)	10 (27%)	9 (25%)
15+ years	14 (38%)	13 (34%)	13 (36%)
Continent			
Europe	23 (62%)	26 (68%)	25 (69%)
North America	14 (38%)	12 (32%)	11 (31%)
Total	37	38	36

regarding adjustments based on probability and nonprobability schemes, auxiliary variables, and simple demographics (questionnaire available upon request). The first questionnaire was sent in early November 2014, and data were collected until late November 2014.

The second wave was based on the responses of the first round by narrowing down the ideas and opinions presented in the first wave. This was done through a systematic content analysis as described by Neuman (2011). We analyzed responses and grouped them by themes, which were used as a guide for the questions asked in the second wave. The second wave questionnaire consisted of 39 questions which pertained to weighting in samples based on no probability scheme, auxiliary variables, and simple demographics (questionnaire available upon request). The questions from the first wave regarding weighting techniques in probability samples were not included in the second wave as they did not provide any new or interesting piece of information. We sent the second questionnaire in February 2015 and data were collected until the beginning of March 2015.

The third wave was a modification of the second wave in that experts were presented with their individual responses as well as group responses from the second round (questionnaire available upon request). The answers that the experts have chosen in Wave 2 were prefilled in the final round so experts could see where their opinions and practices lined up with group responses. We asked the experts to re-evaluate their initial responses from the second wave. The experts were able to change their responses in light of the group response or keep their original responses, with the aim of creating a consensus among the experts. The third wave comprised 10 questions. It was sent at the end of March 2015, with responses collected until early April 2015.

4. Results

4.1. *Opinions of Delphi experts*

Figure 1 shows that 20 of 36 experts agree that users of survey data should correct biases altogether instead of step by step. The alleged reasons are that different errors are intermingled, that auxiliary variables are not specific to one bias, that correcting biases separately complicates the adjustment, that adjusting for biases separately inflates standard errors, and that the different biases are correlated to one another. The other 12 experts advocate for the separate correction of biases. They argue that some auxiliary variables relate to specific errors.

Nine of the 20 experts favor calibration as the best technique and six favor adjustment by propensity score, which may be more practical, as one expert explains. Three experts recommend combining calibration and adjustment by propensity score. For seven experts, the best technique for reducing combined bias is poststratification; for four experts, estimation by raking ratio; and for two experts, regression. Ten of 12 experts who say that biases should be corrected separately also favor a combination of calibration and adjustment by propensity score and using propensity scores as weights in the calibration. A single expert prefers calibration alone but did not explain why. None of the experts in favor of correcting biases separately chose adjustment

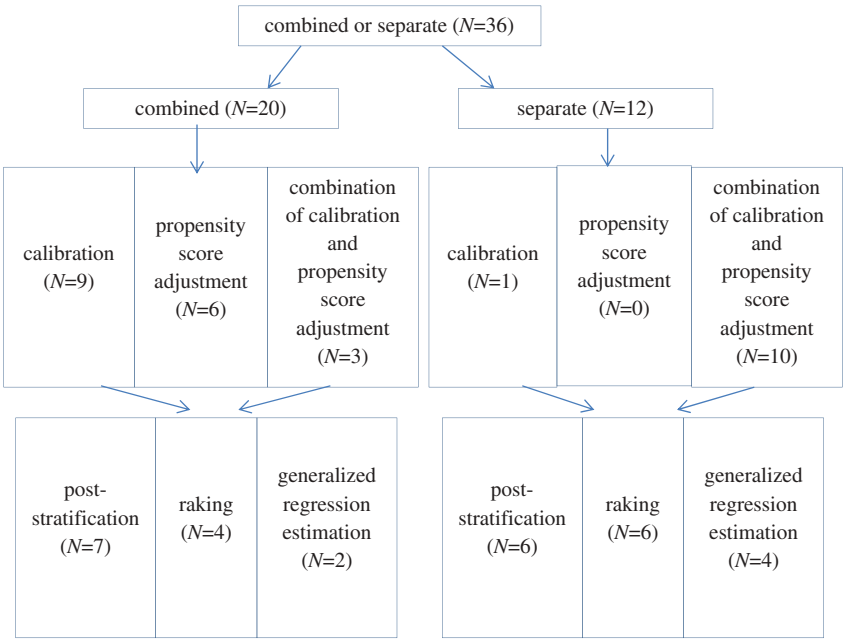

Figure 1. Experts' opinion on the most efficient techniques for correcting biases in web surveys. *N* is the total number of respondents. Arrows indicate routing in the questionnaire. *N* varies from stratum to stratum due to nonresponse.

by propensity score, six of them chose poststratification as the best calibration method to use for separate bias adjustment, and the other six chose estimation by raking ratio, four of them advised estimation by regression as the best calibration method to reduce biases separately.

4.2. *Practices of the Delphi panel*

Twenty-two experts had evaluated weights in their own research, 15 had corrected biases altogether, and 7 separately (Figure 2). Seven experts use calibration methods to correct combined biases, three use propensity score adjustments, and four use a combination of both calibration and adjustments by propensity score. Seven experts use poststratification and seven use estimation by raking ratio to correct biases altogether. The remaining one expert uses regression in his adjustments. Four experts mix calibration methods and adjustments by propensity score to correct biases separately, and two use calibration only. None employ adjustment by propensity only. Four experts use poststratification and four use estimation by raking ratio. One expert uses generalized regression for calibration. The experts who correct a combined bias declare themselves more satisfied with their adjustments than those who correct biases separately.

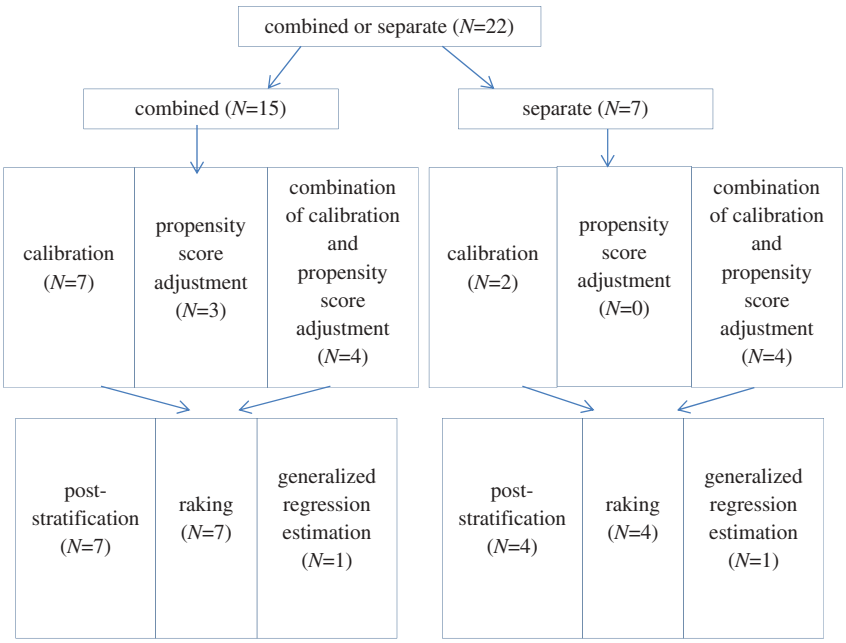

Figure 2. Actual use by the experts of the techniques used in correcting web survey biases. *N* is the total number of respondents. Arrows indicate routing in the questionnaire. *N* varies from stratum to stratum due to nonresponse.

4.3. *Auxiliary variables*

Thirty-one of 36 favor using both socio-demographic and web-related variables because a given outcome involves both types. One expert retains age, sex, education, income, and Internet use or web savviness as the best auxiliary variables. Four experts prefer to use socio-demographics alone. One respondent mentioned that the most suitable auxiliary variables depend on the topic; for surveys representative of the general public, census variables are better; and for private studies (such as association membership), basic characteristics, such as profession, age, and education, are better. None of the experts thought that web-related variables work better on their own. Table 2 shows that contrary to what the experts claimed, only four experts actually apply both socio-demographic and web-related variables in their adjustments; 15 use socio-demographic auxiliary variables. Only one expert uses web-related variables alone.

At the third wave, the experts knew all responses at wave 2. We fixed 7 out of 10 experts to be the majority vote, which corresponds to a score of 3.5 on a scale of 1 to 5. Table 3 presents all auxiliary variables from the highest to the lowest score. Age, education, voting behavior in past election, sex, attitudes toward surveys, attitudes toward the importance of surveys to society, and frequency of Internet access reached 3.5 or higher. These variables obtained agreement from 70–82% among the experts. The other auxiliary variables range from 50–69% of agreement, which remains a high enough score.

4.4. *Differences among experts*

We found no significant difference between European and North American experts, nor among affiliations (market research, government, and academia).

5. Conclusion

We inquired how experts use weighting techniques in web surveys based on no probability schemes. One question is whether survey experts correct for biases together or separately. We found that two out of three experts correct

Table 2. Experts' opinions and practices on auxiliary variables the most efficient at reducing biases.

Types of auxiliary variables	Opinion	Practice
Socio-demographic	4 (11%)	15 (75%)
Web-related	0 (0%)	1 (5%)
Both sociodemographic and web-related	31 (89%)	4 (20%)
Total number of responses	35	20

Table 3. Consensus on auxiliary variables (from the most to the least efficient).

Auxiliary variables used in the weighting scheme	Adjustment rating on a 5-point scale	Percent of agreement
Age	4.1	82*
Education	4.1	81*
Voting behavior in past election	3.8	76*
Sex	3.6	72*
Attitudes toward surveys	3.6	71*
Attitudes toward the importance of surveys	3.6	71*
Access to internet or frequency of internet use	3.5	70*
Ethnic background	3.4	68
Mobility (transportation)	3.4	67
Income	3.3	66
Satisfaction with own health	3.2	65
Voting intention	3.2	65
Occupation (type, sector, part or full time)	3.2	64
Political activism	3.2	64
Interest in the news	3.2	64
Attitudes toward politics, religion, or government	3.1	62
Attitudes toward internet	3.0	61
Household variables (household size, total number of rooms, household income)	3.0	61
Region or district	3.0	60
Customer satisfaction	2.9	57
Leisure activities (theater, museums, restaurants, sports, travel)	2.6	52
Marital status	2.6	52
Membership of organizations	2.5	51
Visits to particular websites	2.5	51
Tastes in music, food, reading	2.5	50

*Consensus defined as agreement over 70%.

for biases altogether, and those who do are more satisfied with adjustments than those who correct biases separately.

We found most experts who correct combined biases use calibration techniques, and most experts who correct biases separately use a combination of adjustment by calibration and propensity scores. Because adjustment by propensity score are used in weighting adjustments for nonresponse (Kalton and Flores-Cervantes, 2003) and because most experts using adjustment by propensity score correct biases separately, experts first account for nonresponse in the propensity scores and then correct coverage or self-selection through calibration. Experts use the same calibration technique for biases combined or separately. Consequently, experts favor poststratification and raking as calibration methods, no matter if they correct biases together or separately. They chose the best method based on available information. For example, the knowledge of the population joint distribution leads them to use poststratification and only partial knowledge leads them to use raking. Experts prefer poststratification and raking over regression because they lack population parameters.

Our purpose was to identify auxiliary variables used in web survey adjustments and which ones are the most efficient. Even though the experts agree that a

combination of socio-demographic and web-related variables is the most efficient in reducing biases, most experts actually use sociodemographic variables.

The gap between wish and practice is due to the availability of auxiliary variables. One expert confirmed: "It is not as if we all have five tools and choose one; it is more like what can be done in a given situation." The point is to collect web-related variables so that they are available for postsurvey adjustments. Because information is seldom available on these variables, an additional probability survey conducted without selection or measurement error can provide these estimates.

The experts agree that the most efficient variables in reducing bias are: age, education, voting behavior in the past election, sex, attitude toward surveys, attitudes toward the importance of surveys to society, and Internet usage. Age, education, and sex are often easily available. Voting behavior, attitudes toward surveys, and attitudes toward the importance of surveys are more difficult to collect, but Couper (1997) shows that these variables, especially in the presence of negative attitudes, are related to both nonresponse and outcome variables. A variable measuring internet use serves as an indicator for self-selection into a web survey. Experts agree that all the auxiliary variables have the same power to reduce biases; efficiency is first a matter of the number of auxiliary variables before being a matter of techniques.

We found a quasi-consensus among experts to apply the methods. Europe or North America, academia, government, or market research made no difference. Assessing the efficiency of weighting techniques and auxiliary variables remains subjective. In the absence of population estimates, there is no way to assess the quality of the adjustment. This is a direction for future research, in the hope of taking advantage of quick and inexpensive web surveys based on no probability schemes compared to surveys based on probability schemes.

Acknowledgments

We thank Peter Lugtig and Barry Schouten for helpful comments and the experts for their participation, among whom Seppo Laaksonen, Hans Walter Steinhauer, Peter Mohler, Brady West, Harm Hartman, and Jannes Hartkamp.

References

Baker, R., Brick, J., Bates, N., Battaglia, M., Couper, M., Dever, J., and Tourangeau,R. (2013). *Report of the AAPOR task force on non-probability sampling* (Technological Report), American Association for Public Opinion Research. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Bethlehem, J. and Biffignandi, S. (2012). *Handbook of Web Surveys*. Hoboken, NJ: Wiley & Sons.Couper, M. (1997). Survey introductions and data quality. *Public Opinion Quarterly*, *61*(2): 317–338.

Delbecq, A., Ven (de), A. V., and Gustafson, D. (1975). *Group Techniques for ProgramPlanning: A Guide to Nominal and Delphi Processes*. Glenview, IL: Scott, Foresman and Co.

Deming, W. and Stephan, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4): 427–444.

Dever, J., Rafferty, A., and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, *2*(2): 47–62.

Deville, J. and Sarndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418): 376–382.

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.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2): 81–97.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, *22*(2): 329–349.

Loosveldt, G. and Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, *2*(2): 93–105.

Malhorta, N. and Krosnick, J. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 anes to Internet surveys with non-probability samples. *Political Analysis*, *15*(3): 286–323.

Neuman, W. L. (2011). *Social Research Methods: Qualitative and Quantitative Approaches*. Boston: Pearson Education.

Okoli, C. and Pawlowski, S. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management*, *42*(1): 15–29.

Pedrazza (de), P., Tijdens, K., and Bustillo (de), R. (2007). Sample bias, weights and efficiency of weights in a continuous web voluntary survey. *ASIS Working Paper 2007-60*, Amsterdam: University of Amsterdam.

Powel, C. (2003). Early indicators of child abuse and neglect: A multi-professional Delphi study. *Wiley InterScience*, *12*(1): 25–40.

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

Scherpenzeel, A. and Bethlehem, J. (2010). How representative are online panels? Problems of coverage and selection and possible solutions. In M. Das, P. Ester, L. Kaczmirek, *Social and Behavioral Research and the Internet*. Oxford: Routledge, 105–130.

Schonlau, M., Soest, A. V., and Kapteyn, A. (2007). Are "webographic" or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods*, *1*(3): 155–163.

Steinmetz, S., Bianchi, A., Tijdens, K., and Biffignandi, S. (2014). Improving web survey quality. In M. Callegaro, R. Baker, J. Bethlehem, A. Gortiz, J. Krosnick, P. Lavrakas, *Online Panel Research: A Data Quality Perspective*. Hoboken, NJ: John Wiley & Sons, 273–310.

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