Selection bias in web surveys

Bethlehem, J.

DOI
10.1111/j.1751-5823.2010.00112.x
Publication date
2010
Document Version
Final published version
Published in
International Statistical Review

International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2010), 78, 2, 161–188 doi:10.1111/j.1751-5823.2010.00112.x

Selection Bias in Web Surveys

Jelke Bethlehem

Statistics Netherlands, Methodology Department, The Hague, The Netherlands
E-mail: jbtm@cbs.nl

Summary

At first sight, web surveys seem to be an interesting and attractive means of data collection. They provide simple, cheap, and fast access to a large group of potential respondents. However, web surveys are not without methodological problems. Specific groups in the populations are under-represented because they have less access to Internet. Furthermore, recruitment of respondents is often based on self-selection. Both under-coverage and self-selection may lead to biased estimates. This paper describes these methodological problems. It also explores the effect of various correction techniques (adjustment weighting and use of reference surveys). This all leads to the question whether properly design web surveys can be used for data collection. The paper attempts to answer this question. It concludes that under-coverage problems may solve itself in the future, but that self-selection leads to unreliable survey outcomes.

Key words: Adjustment weighting; bias; online survey; reference survey; self-selection; under-coverage; web survey.

1 Online Research

1.1 Trends in Data Collection

The survey research landscape has undergone radical changes over the last decades. First, there was the change from traditional paper and pencil interviewing (PAPI) to computer-assisted interviewing (CAI). Couper *et al.* (1998) give an overview of this development. And now, particularly in commercial market research, face-to-face surveys (CAPI), telephone surveys (CATI), and mail surveys (CASI, CASAQ) are increasingly replaced by web surveys. The popularity of online research is not surprising. A web survey is a simple means of getting access to a large group of potential respondents. Questionnaires can be distributed at very low costs. No interviewers are needed, and there are no mailing and printing costs. Surveys can be launched very quickly. Little time is lost between the moment the questionnaire is ready and the start of the fieldwork. Web surveys also offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation, and movies). This raises the question whether web surveys can and should be used in data collection for general population surveys. This paper attempts to answer this question, where the focus is on general population surveys among persons and households.

At first sight, web surveys seem to have much in common with other types of surveys. It is just another mode of data collection. Questions are not asked face-to-face or by telephone, but over the Internet. There are, however, two phenomena that can make the outcomes of web

surveys unreliable: under-coverage and self-selection. They often lead to biased estimates, and therefore wrong conclusions are drawn from the collected data.

Under-coverage means that the sample selection mechanism of the survey is not able to select some elements of the target population. If data is collected by means of the Internet, only respondents with Internet access can complete the questionnaire form. The target population of a survey is, however, usually wider than just those with Internet. This means that people without Internet are excluded from the survey. Research shows that people with Internet access differ, on average, from those without Internet access. As a consequence, web survey results only apply to the sub-population of people having Internet. These results cannot be used to say something about the target population as a whole. Or, to say it differently, web survey based estimates of population characteristics are biased.

Self-selection means that it is completely left to individuals to select themselves for the survey. In case of web surveys, the survey questionnaire is simply put on the web. Respondents are those individuals who happen to have Internet, visit the website and decide to participate in the survey. The survey researcher is not in control over the selection process.

Application of self-selection implies that the principles of probability sampling are not followed. By selecting a random sample, probability theory can be applied, making it possible to construct unbiased estimates. Also the accuracy of estimates can be computed. The probability sampling paradigm has been successfully applied in official and academic statistics since the 1940s, and to a much lesser extent also in more commercial market research. Unfortunately, many web surveys rely on self-selection of respondents instead of probability sampling. This has serious impact on the quality of survey results. The theory of probability sampling cannot be applied and estimates are often substantially biased.

This paper can be seen as a tutorial giving an overview of the methodological aspects of web surveys. It shows how the statistical properties of estimators are affected by under-coverage and self-selection.

It starts by describing how the probability sampling paradigm became the fundament of survey sampling. This is followed of an overview of what can go wrong in surveys. Two of these problems are particularly relevant for web surveys: under-coverage and self-selection. These problems are discussed in more detail. By taking traditional sampling theory as a starting point, and adding some elements from non-response theory, expressions are obtained for the bias due to under-coverage and self-selection.

A bias due to nonresponse can sometimes be reduced by applying correction techniques such as adjustment weighting. It is explored whether these techniques can also be helpful in reducing the bias due to under-coverage or self-selection. Particularly, attention is paid to using a reference survey as a means to correct web survey results.

A number of simulation experiments are described that show the effects of under-coverage and self-selection. They confirm conclusions earlier in the paper that correction techniques need not always be successful. Moreover, use of reference surveys may dramatically reduce the effective sample size.

1.2 The Probability Sampling Paradigm

The principles of probability sampling are fundamental for modern survey taking. The first ideas only emerged a little more than a century ago. For an overview, see for example Bethlehem (2009b). The idea of compiling statistical overviews of the state of affairs in a country is already very old. Kendall (1960) gives a historical overview. As far back as Babylonian times censuses of agriculture were taken. Ancient China counted its people to determine the revenues and the military strength of its provinces. There are also accounts of statistical overviews compiled by

Egyptian rulers long before Christ. Rome regularly took a census of people and of property. The data were used to establish the political status of citizens and to assess their military and tax obligations to the state.

Censuses were rare in the Middle Ages. The most famous one was the census of England taken by the order of William the Conqueror, King of England. The compilation of this *Domesday Book* started in the year 1086. The book records a wealth of information about each manor and each village in the country. Another interesting example can be found in the Inca Empire that existed between 1,000 and 1,500 in South America. Each Inca tribe had its own statistician, called the *Quipucamayoc*. This man kept records of, for example, the number of people, the number of houses, the number of llamas, the number of marriages, and the number of young men that could be recruited for the army. All these facts were recorded on a *quipu*, a system of knots in coloured ropes. A decimal system was used for this.

The development of modern sampling theory started around the year 1895. In that year, Anders Kiaer (1895, 1997), the founder and first director of Statistics Norway, published his *Representative Method*. It was a partial inquiry in which a large number of persons were questioned. This selection formed a "miniature" of the population. Persons were selected arbitrary, but according to some rational scheme based on general results of previous investigations. Anders Kiaer stressed the importance of *representativity*. His argument was that, if a sample was representative with respect to variables for which the population distribution was known, it would also be representative with respect to the other survey variables.

A basic problem of the Representative Method was that there was no way of establishing the accuracy of estimates. The method lacked a formal theory of inference. It was Bowley (1906, 1926), who made the first steps in this direction. He showed that for large samples, selected at random from the population with equal probabilities, estimators had an approximately normal distribution.

From this moment on, there were two methods of sample selection. The first one was Kiaer's Representative Method, based on purposive selection, in which representativity played a crucial role, and for which no measure of the accuracy of the estimates could be obtained. The second was Bowley's approach, based on simple random sampling, and for which an indication of the accuracy of estimates could be computed. Both methods existed side by side for a number of years. This situation lasted until 1934, in which year the Polish scientist Jerzy Neyman published his now famous paper, see Neyman (1934). Neyman developed a new theory based on the concept of the confidence interval. By using random selection instead of purposive selection, there was no need any more to make prior assumptions about the population. Neyman also showed that the Representative Method based on purposive sampling failed to provide satisfactory estimates of population characteristics. As a result, the method of purposive sampling fell into disrepute in official statistics.

The classical theory of survey sampling was more or less completed in 1952. Horvitz & Thompson (1952) developed a general theory for constructing unbiased estimates. Whatever the selection probabilities are, as long as they are known and positive, it is always possible to construct a reliable estimate. The probability sampling approach was almost unanimously accepted. Most of the classical books about sampling were also published by then: Yates (1949), Deming (1950), Cochran (1953), and Hansen *et al.* (1953).

1.3 About Errors in Surveys

One of the main objectives of a sample survey is to compute estimates of population characteristics. Such estimates will never be exactly equal to the population characteristics. There will always be some error. This error can have many causes. Bethlehem (2009a) presents

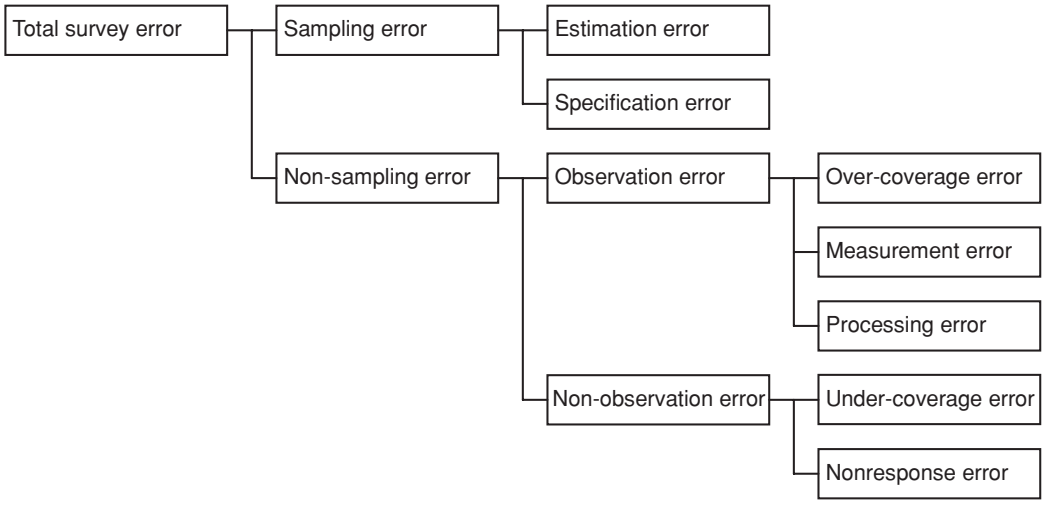

Figure 1. *Taxonomy of survey errors.*

a taxonomy of possible causes. A slightly adapated version is presented in Figure 1. The taxonomy is a more extended version of the one given by Kish (1967).

The ultimate result of all these errors is a discrepancy between the survey estimate and the population characteristic to be estimated. This discrepancy is called the *total survey error*. Two broad categories can be distinguished contributing to this total error: sampling errors and non-sampling errors.

Sampling errors are introduced by the sampling design. They are due to the fact that estimates are based on a sample and not on a complete enumeration of the population. Sampling errors vanish if the complete population is observed. Since only a sample is available for computing population characteristics, and not the complete data set, one has to rely on estimates. The sampling error can be split in a estimation error and a specification error.

The *estimation error* denotes the effect caused by using a sample based on a random selection procedure. Every new selection of a sample will result in different elements, and thus in a different value of the estimator. The estimation error can by quantified by applying probability theory. Its magnitude can be controlled through the sampling design. For example, by increasing the sample size, or by taking selection probabilities proportional to the values of some well chosen auxiliary variable, one can reduce the estimation error.

The estimation error can have a different effect in a web surveys depending on the sampling mechanism used. If a proper sample is selected from a sampling frame that is independent of the Internet (e.g. a population register), the effect is the same as for other types of surveys. If sampling comes down to self-selection of respondents, there is also an estimation error, but it cannot be quantified since selection probabilities are unknown.

A *specification error* occurs when the true selection probabilities differ from the selection probabilities specified in the sampling design. This happens, for example, when elements have multiple occurrences in the sampling frame without this being known. Selection errors are hard to avoid without thorough investigation of the sampling frame.

A specification error may occur in a web survey that is selected from a sampling frame. In the case of self-selection the selection probabilities are even unknown, so there are no anticipated probabilities (unless a naive researcher assumes all selection probabilities to be the same).

Non-sampling errors may even occur if the whole population is investigated. They denote errors made during the process of obtaining answers to questions asked. Non-sampling errors can be divided in observation errors and non-observation errors.

Observation errors are one form of non-sampling errors. They refer to errors made during the process of obtaining and recording answers. An *over-coverage error* means that elements are included in the survey and that do not belong to the target population. A *measurement error* occurs when a respondent does not understand a question, or does not want to give the true answer, or if the interviewer makes an error in recording the answer. Also, interviewer effects, question wording effects, and memory effects belong to this type of error. A measurement error causes a difference between the true value and the value processed in the survey. A *processing* error denotes an error made during data processing, for example data entry.

Over-coverage errors and measurement errors can occur in web survey even if a proper sampling frame is used. Over-coverage can even be a more serious problem if the web survey is based on self-selection. Then there is no control at all over who completes the questionnaire. A web survey is a self-administered survey. There are no interviewers. Therefore, the design of the questionnaire form is of vital importance. Dillman *et al.* (2008) and Couper (2008) show that design flaws can lead to serious measurement errors.

Non-observation errors are errors made because intended measurements cannot be carried out. *Under-coverage* occurs when elements of the target population do not have a corresponding entry in the sampling frame. These elements can and will never be contacted. Another non-observation error is *non-response*. It is the phenomenon that elements selected in the sample do not provide the required information.

Under-coverage is a serious problem if the Internet is used for data collection and the target population is wider than the Internet population. Web surveys also suffer from non-response. A web survey questionnaire is a form of self-administered questionnaire. Therefore, web surveys have a potential of high non-response rates. An additional source of non-response are technical problems of respondents having to interact with the Internet, see for example Couper (2000), Dillman & Bowker (2001), Fricker & Schonlau (2002), and Heerwegh & Loosveldt (2002). Slow modem speeds, unreliable connections, high connection costs, low-end browsers, and unclear navigation instructions may frustrate respondents. This often results in respondents discontinuing the completion of the questionnaire. In order to keep the survey response up to an acceptable level, every measure must be taken to avoid these problems. This requires a careful design of web survey questionnaire instruments.

This overview makes clear that a lot can go wrong during a survey, and usually it does. Some errors can be avoided by taking preventive measures at the design stage. However, some errors will remain. The same applies to web surveys, and some problems are even more severe for web surveys.

1.4 Coverage Problems of Web Surveys

Web surveys may suffer from serious coverage problems, whether the sample is selected by means of probability sampling or self-selection. The reason is that people without Internet access will never be able to participate in a web survey. If those with Internet differ from those without Internet, survey results will be biased. The coverage problem is described in more detail in this section.

The collection of all elements that can be contacted through the sampling frame is called the *frame population*. Since the sample is selected from the frame population, conclusions drawn from the collected survey data will apply to the frame population, and not necessarily to the

Table 1
Internet access by households in Europe in 2007. Source: Eurostat

Rank	Country	Internet access	Broadband connection
1	Netherlands	83%	74%
2	Sweden	79%	67%
3	Denmark	78%	70%
4	Luxembourg	75%	58%
5	Germany	71%	50%
...			
23	Hungary	38%	33%
24	Czech Rep.	35%	28%
25	Greece	25%	7%
26	Romania	22%	8%
27	Bulgaria	19%	15%
	EU	54%	42%

target population. Coverage problems arise when the frame population differs from the target population.

Under-coverage occurs when elements in the target population do not appear in the frame population. These elements have zero probability of being selected in the sample. Under-coverage can be a serious problem for Internet surveys. If the target population consists of all people with an Internet connection, there is no problem. However, the target population can be wider than that, particularly for general population surveys. Then, under-coverage occurs due to the fact that still many people do not have access to the Internet. According to Eurostat (2007), the statistical office of the European Union, countries in the Union differ substantially in Internet coverage of households. Table 1 summarizes the extremes. The figures are survey-based estimates, and therefore may be subject to errors. For example, the Dutch percentages come from a CAPI survey conducted by Statistics Netherlands.

Internet access is very high in The Netherlands. More than four out of five households have an Internet connection. Internet coverage is also high in the Scandinavian countries Sweden and Denmark. Coverage is very low in the Balkan countries Romania and Bulgaria. Only approximately one out of five households there has Internet access.

Table 1 also contains information about the percentage of households having a broadband Internet connection. It is clear that still many Internet connections are based on slow modems. This may put restrictions on the questionnaires used. They may not be too long and too complicated, and prohibit the use of advanced features like images, video and animation. Slow questionnaire processing may cause respondents to break off the session, resulting in only partially completed questionnaires.

In the Netherlands, the percentage of households having an Internet connection increased in six years time from 63% to 86%, see Figure 2 (source: CBS, 2008). Still, it is clear that not every household will have access to Internet in the near future.

Analysis of data with respect to Internet access in The Netherlands in 2005 shows that Internet access is unevenly distributed over the population. These data have been collected in the Integrated Survey on Household Living Conditions. This survey is based on a probability sample from the Dutch population. Each person had an equal probability of selection. The mode of data collection was face-to-face interviewing (CAPI). There were no under-coverage problems in this survey. The response rate was around 65%. A comprehensive weighting adjustment procedure was carried out to correct for a possible bias due to nonresponse.

The results show that more males than females have access to the Internet. The percentage of males with Internet is 86%, whereas for females it is only 81%.

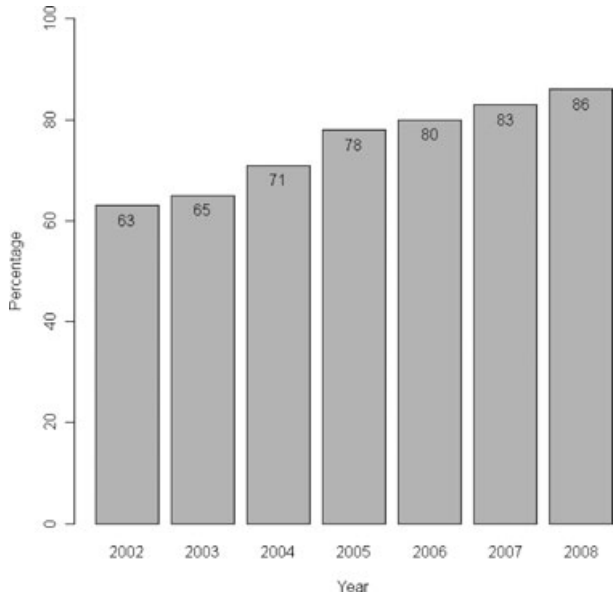

Figure 2. *Percentage of persons having Internet.*

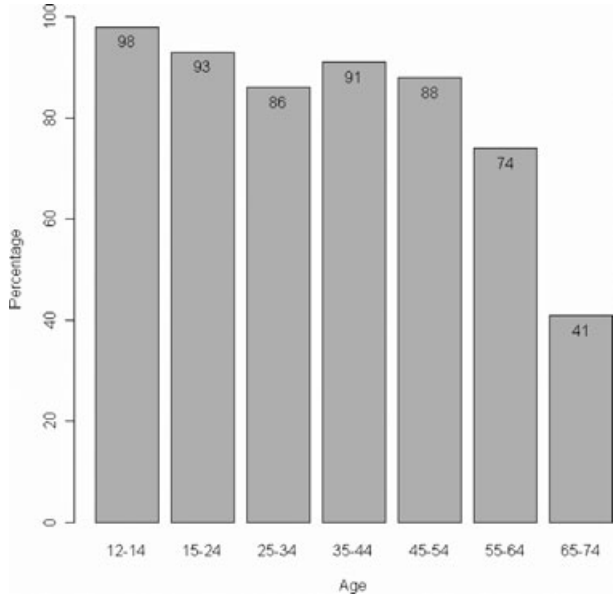

Figure 3. *Internet access by age.*

Figure 3 contains the percentage of people having Internet by age group (in 2005). Internet access at home decreases with age. Particularly, people of age 55 and older will be very much under-represented when the Internet is used for data collection.

Figure 4 contains Internet access by level of education (in 2005). It is clear that people with a higher level of education more frequently have Internet than people with a lower level of education.

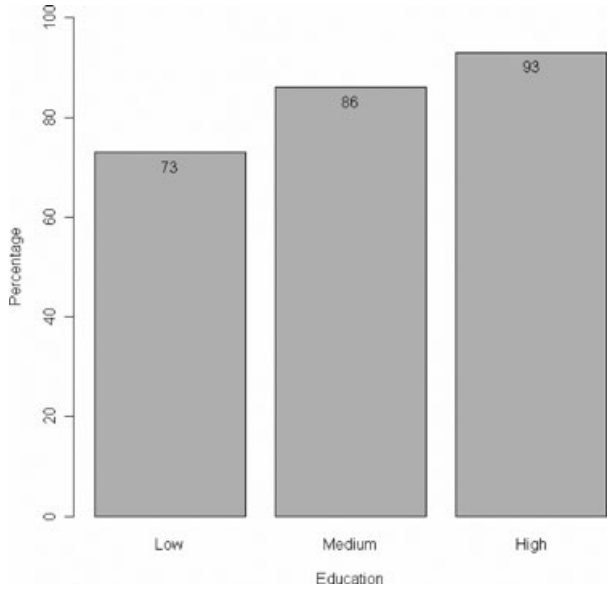

Figure 4. *Internet access by level of education.*

According to De Haan & Van't Hof (2006) Internet access among non-native young people is much lower in The Netherlands than among native young people: 91% of the young natives have access to Internet. This is 80% for young people from Surinam and Antilles, 68% for young people from Turkey and only 64% for young people from Morocco. These observations are in line with the findings of authors in other countries. See for example Couper (2000), and Dillman & Bowker (2001).

It is clear that use of the Internet for data collection will cause problems, because certain specific groups are substantially under-represented. Even if a proper probability sample is selected, the result will be a selective sample. Specific groups in the target population will not be able to fill in the (electronic) questionnaire form.

Note that there is some similarity with CATI surveys in which only telephone directories are used as a sampling frame. Here, people without a phone or with an unlisted number will be excluded from the survey.

1.5 Selection Problems in Web Surveys

Horvitz & Thompson (1952) show in their seminal paper that unbiased estimates of population characteristics can be computed only if a real probability sample has been used, every element in the population has a non-zero probability of selection, and all these probabilities are known to the researcher. Furthermore, only under these conditions, the accuracy of estimates can be computed.

Web surveys appear in many different forms, from simple e-mail surveys to professionally designed interactive forms. There are web surveys based on probability sampling, for example surveys among students of a university or employees of a large company. Unfortunately, many web surveys, particularly those conducted by market research organisations, are not based on probability sampling. The survey questionnaire 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 over the selection process. Therefore, no unbiased estimates can be computed nor can the accuracy of estimates be determined. These surveys are called *self-selection surveys*.

As an illustration, an overview of the situation in The Netherlands shows that all major opinion polls use web panels that have been set up by means of self-selection. The values of some demographic variables are recorded during the recruitment phase. Therefore the distribution of these variables in a poll can be compared with their distribution in the population. Weighting adjustment techniques can be applied in an attempt to correct for over- or under-representation of specific groups. An example of large online cross-sectional survey in The Netherlands was *21minuten.nl*, a survey supposed to supply answers to questions about important problems in Dutch society. The first edition of this survey was conducted in 2006. Within a period of six weeks about 170,000 people completed the online questionnaire. A similar survey was conducted in Germany (*Perspektive Deutschland*). Vonk *et al.* (2006) describe a study across 19 online panels of Dutch market research organizations. It shows that most of them use self-selection. Up until now, self-selection web surveys have not been used by governmental survey organizations in The Netherlands.

Self-selection web surveys results are sometimes claimed to be "representative" because of the high number of respondents or as a results of advanced adjustment weighting procedures. The term representative is rather confusing. Kruskal & Mosteller (1979a, 1979b, 1979c) show that it can have many meanings and it is often used in a very loose sense to convey a vague idea of good quality. It is even sometimes claimed that a large number of respondents ensures validity and reliability. Unfortunately, it is a well-known fact in the survey methodology literature that this is not the case. It is shown again in this paper.

As an example, the effects of self-selection are illustrated using survey results related to the general elections in The Netherlands in 2006. Various market research organizations carried out opinion polls in an attempt to predict the outcome of these elections. The results of the three major polls are summarized in Table 2. The table contains numbers of seats. There is a direct linear relationship between numbers of seats and percentages of votes.

Politieke Barometer, *Peil.nl*, and *De Stemming* are opinion polls based on samples from online panels (also called access panels). To reduce a possible bias, adjustment weighting has been carried out. *DPES* is the Dutch Parliamentary Election Study. The fieldwork for this comprehensive survey was carried out by Statistics Netherlands. It used a true (two-stage) probability sample. Respondents were interviewed face-to-face (using CAPI). Underscored

Table 2
Dutch Parliamentary elections 2006. Official results and predictions of various opinion surveys.

	Election result	Politieke Barometer	Peil.nl	De Stemming	DPES 2006
Sample size		1,000	2,500	2,000	2,600
Seats in parliament					
CDA (Christian democrats)	41	41	42	41	41
PvdA (Social democrats)	33	37	38	31	32
VVD (Liberals)	22	23	22	21	22
SP (Socialists)	25	23	23	32	26
GL (Green party)	7	7	8	5	7
D66 (Liberal democrats)	3	3	2	1	3
ChristenUnie (Christan)	6	6	6	8	6
SGP (Christian)	2	2	2	1	2
PvdD (Animal party)	2	2	1	2	2
PvdV (Conservative)	9	4	5	6	8
Other parties	0	2	1	2	1
Mean absolute difference		1.27	1.45	2.00	0.36

numbers denote predictions differing three seats or more from the true result. This happened only for the web panels, and not for the election survey. These predictions were even considered unsatisfactory by the organizations that produced them. It is clear that in this example the *DPES* outperformed the online polls.

Probability sampling has the additional advantage that it provides protection against certain groups in the population attempting to manipulate the outcomes of the survey. This may typically play a role in opinion polls. Self-selection does not have this safeguard. An example of this effect could be observed in the election of the 2005 Book of the Year award (Dutch: NS Publieksprijs), a high-profile literary prize. The winning book was determined by means of a poll on a website. People could vote for one of the nominated books or mention another book of their choice. More than 90,000 people participated in the survey. The winner turned out to be the new interconfessional Bible translation launched by the Netherlands and Flanders Bible Societies. This book was not nominated, but nevertheless an overwhelming majority (72%) voted for it. This was due to a campaign launched by (among others) Bible societies, a Christian broadcaster and Christian newspaper. Although this was all completely within the rules of the contest, the group of voters could clearly not considered to be representative for the Dutch population.

2 Sampling the Internet Population

2.1 Basic Theoretical Concepts

This section provides a theoretical framework for sampling the Internet-population. After introducing basic notations, expressions for the bias of estimators are derived. The approach is similar to that of analysing the effects of non-response, see for example Bethlehem (1988, 2002).

Let the target population of the survey consist of N identifiable elements, which are labelled $1, 2, \ldots, N$. Associated with each element k is a value Y_k of the target variable Y. The aim of the web survey is assumed to be estimation of the population mean

$$\bar{Y} = \frac{1}{N} \sum_{k=1}^{N} Y_k \tag{1}$$

of the target variable Y.

The population U is divided into two sub-populations U_I of elements having access to Internet, and U_{NI} of elements not having access to the Internet. Associated with each element k is an indicator I_k, where $I_k = 1$ if element k has access to the Internet ($k \in U_I$), and $I_k = 0$ otherwise ($k \in U_{NI}$). The sub-population U_I is called the *Internet population*. Let $N_I = I_1 + I_2 + \cdots + I_N$ denote the size of the Internet population U_I. Likewise, N_{NI} denotes the size of the sub-population without Internet, where $N_I + N_{NI} = N$.

The mean of the target variable for the elements in the Internet population is equal to

$$\bar{Y}_I = \frac{1}{N_I} \sum_{k=1}^{N} I_k Y_k. \tag{2}$$

2.2 A Random Sample from the Internet Population

Suppose, it is possible to draw a simple random sample without replacement from the Internet population. This more or less ideal case requires a sampling frame listing all elements having

access to the Internet. Such a sampling frame is sometimes available, for example for a survey of university students or employees of a company. Unfortunately, no such list exists for general population surveys. One way to solve this problem is to select a random sample from a larger sampling frame (e.g. a population or address register), approach the selected people in a classical way (by mail, telephone, or face-to-face), and filter out only those people having access to the Internet. Selected people are provided with an Internet address where they can fill in the questionnaire form. Use of unique access code guarantees that a sample person can complete the questionnaire only once, and that only sampled persons can participate in the survey. It is clear that such frames suffer from over-coverage, but with this approach every element in the Internet population has a positive and known probability of being selected.

A random sample selected without replacement from the Internet population is represented by a vector

$$s = (s_1, s_2, \ldots, s_N) \tag{3}$$

of N indicators, where the k-th indicator s_k assumes the value 1 if element k is selected, and otherwise it assumes the value 0, for $k = 1, 2, \ldots, N$. Note that always $s_k = 0$ for elements k outside the Internet population. The sample size is denoted by $n_I = s_1 + s_2 + \cdots + s_N$.

The expected value $\pi_k = E(s_k)$ is called the *first order inclusion probability* of element k. Horvitz & Thompson (1952) have shown that always an unbiased estimator of the population mean can be defined if all elements in the population have known, positive first order inclusion probabilities. The Horvitz-Thompson estimator for the mean of the Internet population is defined by

$$\bar{y}_{HT} = \frac{1}{N_I} \sum_{k=1}^{N} s_k I_k \frac{Y_k}{\pi_k}, \tag{4}$$

where by definition $Y_k/\pi_k = 0$ for all elements outside the Internet population. In the case of a simple random sample from the Internet population, all first order inclusion probabilities are equal to n/N_I. Therefore expression (4) reduces to

$$\bar{y}_I = \frac{1}{n} \sum_{k=1}^{N} s_k I_k Y_k. \tag{5}$$

This estimator is an unbiased estimator of the mean $\bar{Y}_I$ of the Internet population, but not necessarily of the mean $\bar{Y}$ of the target population. The bias is equal to

$$B(\bar{y}_{HT}) = E(\bar{y}_{HT}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}). \tag{6}$$

Expression (6) shows that the magnitude of this bias is determined by two factors.

- The proportion N_{NI}/N of people without Internet access. The bias will increase as a larger proportion of the population does not have access to the Internet.
- The *contrast* $\bar{Y}_I - \bar{Y}_{NI}$ between the Internet population and the non-Internet population. The more the mean of the target variable differs for these two sub-populations, the larger the bias will be.

The size of the non-Internet population cannot yet be neglected in most countries. Table 1 shows that the percentage of people without Internet is still substantial in many countries.

Furthermore, there are substantial differences between people with and without Internet. For example, the graphs in Section 1.4 show that, at least in The Netherlands, specific groups are

under-represented in web surveys, for example the elderly, those with a low level of education, and ethnic minority groups. So, the conclusion is that a random sample from an Internet population will often lead to biased estimates for the target population.

2.3 Self-Selection from the Internet Population

Many web surveys rely on self-selection of respondents. Participation requires in the first place that respondents are aware of the existence of a survey (they have to accidentally visit the website, or they have to follow up a banner or an e-mail message). In the second place, they have to make the decision to fill in the questionnaire on the Internet. All this means that each element k in the Internet population has unknown probability ρ_k of participating in the survey, for $k = 1, 2, \ldots, N_I$. The responding elements are denoted by a vector

$$r = (r_1, r_2, \ldots, r_N) \tag{7}$$

of N indicators, where the k-th indicator r_k assumes the value 1 if element k participates, and otherwise it assumes the value 0, for $k = 1, 2, \ldots, N$. The expected value $\rho_k = E(r_k)$ is called the *response probability* of element k. For sake of convenience response probabilities are also introduced for elements in the non-Internet population. By definition the values of all these probabilities are 0.

The realised sample size is denoted by

$$n_S = \sum_{k=1}^{N} r_k. \tag{8}$$

A naive researcher assuming that every element in the Internet population has the same probability of being selected in the sample, will use the sample mean

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^{N} r_k Y_k, \tag{9}$$

as an estimator for the population mean. The expected value of this estimator is approximately equal to

$$E(\bar{y}_S) \approx \bar{Y}_I^* = \frac{1}{N_I \bar{\rho}} \sum_{k=1}^{N} \rho_k I_k Y_k, \tag{10}$$

where $\bar{\rho}$ is the mean of all response propensities in the Internet population. This expression was derived by Bethlehem (1988).

It is clear that, generally, the expected value of the sample mean is not equal to the population mean of the Internet population. One situation in which the bias vanishes is that in which all response probabilities in the Internet population are equal. In terms of nonresponse correction theory, this comes down to Missing Completely Missing At Random (MCAR). This is the situation in which the occurrence of nonresponse is completely independent of all variables measured in the survey. For more information on MCAR and other missing data mechanisms, see Little & Rubin (2002). Indeed, in the case of MCAR self-selection does not lead to an unrepresentative sample because all elements have the same selection probability.

Bethlehem (1988, 2002) shows that the bias of the sample mean (9) can be written as

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y}_I \approx \bar{Y}_I^* - \bar{Y}_I = \frac{C(\rho, Y)}{\bar{\rho}} = \frac{R(\rho, Y) S(\rho) S(Y)}{\bar{\rho}}, \tag{11}$$

in which

$$C(\rho, Y) = \frac{1}{N_I} \sum_{k=1}^{N} I_k(\rho_k - \bar{\rho})(Y_k - \bar{Y}_I) \tag{12}$$

is the covariance between the values of target variable and the response probabilities in the Internet-population and $\bar{\rho}$ is the average response probability. Furthermore, $R(\rho,Y)$ is the correlation coefficient between target variable and the response behaviour, $S(\rho)$ is the standard deviation of the response probabilities and $S(Y)$ is the standard deviation of the target variable. The bias of the sample mean (as an estimator of the mean of the Internet population) is determined by two factors:

- The average response probability. If people are more likely to participate in the survey, the average response probability will be higher, and thus the bias will be smaller.
- The variation in response probabilities. The more these probabilities vary, the larger the bias will be.
- The relationship between the target variable and response behaviour. A strong correlation between the values of the target variable and the response probabilities, will lead to a large bias.

There are three situations in which this bias vanishes:

(1) All response propensities are equal. Again, this is the case in the which the self-selection process can be compared with a simple random sample.
(2) All values of the target variable are equal. This situation is very unlikely to occur in practice. No survey would be necessary. One observation would be sufficient.
(3) There is no relationship between target variable and response behaviour. It means participation does not depend on the value of the target variable.

Expression (11) for the bias of the estimator can be used to compute an upper bound for the bias. Given the mean response probability $\bar{\rho}$, there is a maximum value the standard deviation $S(\rho)$ of the response probabilities cannot exceed:

$$S(\rho) \leq \sqrt{\bar{\rho}(1 - \bar{\rho})}. \tag{13}$$

This implies that in the worst case ($S(\rho)$ assumes it maximum value and the correlation coefficient $R(\rho, Y)$ is equal to either $+1$ or -1), the absolute value of the bias will be equal to

$$|B_{max}| = S(Y)\sqrt{\frac{1}{\bar{\rho}} - 1}. \tag{14}$$

This worst case value of the bias also applies to the situation in which a probability sample has been drawn and subsequently non-response occurs in the fieldwork. Therefore, expression (14) provides a means to compare potential biases in various survey situations.

For example, general population surveys of Statistics Netherlands have response rates of around 70%. This means the absolute maximum bias is equal to $0.65 \times S(Y)$. One of the largest web surveys in the Netherlands was *21minuten.nl*. This survey was supposed to provide answers to questions about important problems in the Dutch society. Within a period of six week in 2006 about 170,000 people completed the questionnaire (which took about 21 minutes). As everyone could participate in the survey, the target population was not defined properly. If it is assumed the target population consists of all Dutch citizens from the age of 18, the average response probability was 170,000/12,800,000 = 0.0133. Hence the absolute maximum bias is equal to

$8.61 \times S(Y)$. The conclusion is that the bias of the large web survey can be a factor 13 larger that the bias of the small probability survey.

It is usually not the objective of general population surveys to estimate the mean of the Internet population, but the mean of the total population. Then the bias of the sample mean is equal to

$$\begin{aligned} B(\bar{y}_S) &= E(\bar{y}_S) - \bar{Y} = E(\bar{y}_S) - \bar{Y}_I + \bar{Y}_I - \bar{Y} \\ &= \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) + \frac{C(\rho, Y)}{\bar{\rho}}. \end{aligned} \tag{15}$$

The bias consists of two terms: a bias caused by interviewing just the Internet population instead of the complete target population (under-coverage bias) and a bias caused by self-selection of respondents in the Internet population (self-selection bias). Theoretically, it is possible that these two biases compensate one another. If people without Internet resemble people with Internet that are less likely to participate, the combined effects will produce a larger bias. Practical experiences suggest that this may often be the case. For example, suppose Y is a variable measuring the intensity of some activity on the Internet (surfing, playing on-line games, being active in social networks). Then a positive correlation between Y and response propensities is not unlikely. Also the mean of Y for the Internet population will be positive while the mean of the non-Internet population will be 0. So, both bias terms have a positive value.

3 Weighting Adjustment

3.1 Why Weighting Adjustment?

Weighting adjustment is a family of techniques that attempt to improve the accuracy of survey estimates by using of auxiliary information. *Auxiliary information* is defined as a set of variables that have been measured in the survey, and for which information on their population distribution (or complete sample distribution) is available. By comparing the response distribution of an auxiliary variable with its population distribution, it can be assessed whether or not the sample is representative for the population (with respect to this variable). If these distributions differ considerably, one must conclude that the sample is selective. To correct this, adjustment weights are computed. Weights are assigned to all records of observed elements. Estimates of population characteristics are then computed by using the weighted values instead of the unweighted values. Weighting adjustment is often used to correct surveys that are affected by non-response. An overview of weighting adjustment can be found in Bethlehem (2002) and Särndal & Lundström (2005).

This section explores the possibility to reduce the bias of web survey estimates. For sake of convenience it is assumed that a simple random sample has been selected from the Internet population. Section 3.2 analyses the effects of post-stratification, where weights are computed using the distribution of auxiliary variables in the complete population. Section 3.3 investigates the situation where the population distribution of auxiliary variables is estimated using data from a small, true probability sample (a so-called *reference survey*). Section 3.4 explores the possibilities of improving estimates by also sampling the non-Internet population. Finally, Section 3.5 discusses propensity weighting, a weighting technique often applied by commercial market research agencies.

3.2 Post-Stratification

Post-stratification is a well-known and often used weighting method, see for example Cochran (1977) or Bethlehem (2002). To carry out post-stratification, one or more qualitative auxiliary

variables are needed. Here, only one such variable is considered. The situation for more variables is not essentially different. Suppose there is an auxiliary variable X having L categories. So it divides the target population into L strata. The strata are denoted by the subsets $U_1, U_2, \ldots, U_L$ of the population U. The number of target population elements in stratum U_h is denoted by N_h, for $h = 1, 2, \ldots, L$. The population size N is equal to $N = N_1 + N_2 + \cdots + N_L$. This is the population information assumed to be available.

Suppose a simple random sample of size n is selected from the Internet population. If n_h denotes the number of sample elements in stratum h, then $n = n_1 + n_2 + \cdots + n_L$. The values of the n_h are the result of a random selection process, so they are random variables. Note that since the sample is selected from the Internet population, only elements in the sub-strata $U_I \cap U_h$ are observed (for $h = 1, 2, \ldots, L$).

Post-stratification assigns identical adjustment weights to all elements in the same stratum. The weight w_k for an element k in stratum h is equal to

$$w_k = \frac{N_h/N}{n_h/n}. \tag{16}$$

The sample mean

$$\bar{y}_I = \frac{1}{n}\sum_{k=1}^{N} s_k I_k Y_k \tag{17}$$

is now replaced by the weighted sample mean

$$\bar{y}_{I,PS} = \frac{1}{n}\sum_{k=1}^{N} s_k w_k I_k Y_k. \tag{18}$$

Substituting the weights and working out this expression leads to the post-stratification estimator

$$\bar{y}_{I,PS} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_I^{(h)} = \sum_{h=1}^{L} W_h \bar{y}_I^{(h)}, \tag{19}$$

where $\bar{y}_I^{(h)}$ is the sample mean in stratum h and $W_h = N_h/N$ is the relative size of stratum h. The expected value of this post-stratification estimator is equal to

$$E(\bar{y}_{I,PS}) = \frac{1}{N}\sum_{h=1}^{L} N_h E\left(\bar{y}_I^{(h)}\right) = \sum_{h=1}^{L} W_h \bar{Y}_I^{(h)} = \tilde{Y}_I, \tag{20}$$

where $\bar{Y}_I^{(h)}$ is the mean of the target variable in stratum h of the Internet population. Generally, this mean will not be equal to the mean $\bar{Y}^{(h)}$ of the target variable in stratum h of the target population. The bias of this estimator is equal to

$$\begin{aligned} B(\bar{y}_{I,PS}) &= E(\bar{y}_{I,PS}) - \bar{Y} = \tilde{Y}_I - \bar{Y} = \sum_{h=1}^{L} W_h \left(\bar{Y}_I^{(h)} - \bar{Y}^{(h)}\right) \\ &= \sum_{h=1}^{L} W_h \frac{N_{NI,h}}{N_h}\left(\bar{Y}_I^{(h)} - \bar{Y}_{NI}^{(h)}\right), \end{aligned} \tag{21}$$

where $N_{NI,h}$ is the number of elements in stratum h of the non-Internet population.

The bias will be small if there is (on average) no difference between elements with and without Internet within the strata. This is the case if there is a strong relationship between the target

variable Y and the stratification variable X. The variation in the values of Y manifests itself between strata but not within strata. In other words, the strata are homogeneous with respect to the target variable. In non-response correction terminology, this situation comes down to Missing At Random (MAR).

The conclusion is that application of post-stratification will successfully reduce the bias of the estimator if proper auxiliary variables can be found. Such variables should satisfy three conditions:

- They have to be measured in the survey (or complete sample).
- Their population distribution $(N_1, N_2, \ldots, N_L)$ must be known.
- They must be strongly correlated with all target variables.

Unfortunately, such variables are not very often available, or there is only a weak correlation.

It can be shown that, in general, the variance of the post-stratification estimator is equal to

$$V(\bar{y}_{PS}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}^{(h)}). \tag{22}$$

Cochran (1977) shows that in the case of a simple random sampling from the complete population, this expression is equal to

$$V(\bar{y}_{PS}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^{L} (1 - W_h) S_h^2, \tag{23}$$

where $f = n/N$ and S_h^2 is the variance in stratum h. If the strata are homogeneous with respect to Y, the variance of estimator will be small. In the case of a simple random sample from the Internet population, the variance of the estimator (19) becomes

$$V(\bar{y}_{I,PS}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n W_{I,h}} + \frac{1 - W_{I,h}}{(n W_{I,h})^2} - \frac{1}{N_{I,h}} \right) S_{I,h}^2, \tag{24}$$

where $N_{I,h}$ is the size of stratum h in the Internet population, $W_{I,h} = N_{I,h}/N_I$ and $S_{I,h}^2$ is the variance in stratum h of the Internet population.

Post-stratification is only a simple and straightforward weighting technique. More advances techniques are described in Bethlehem (2002) and Särndal & Lundström (2005). It should be noted that all these weighting techniques will only be successful if MAR applies.

3.3 Weighting Adjustment with a Reference Sample

The previous section showed that post-stratification can be an effective correction technique provided auxiliary variables are available that have a strong correlation with the target variables of the survey. If such variables are not available, one might consider conducting a *reference survey*. This reference survey is a (usually small) probability sample, where data collection takes place with a mode different from the web, for example Computer Assisted Personal Interviewing (CAPI, with laptops) or Computer Assisted Telephone Interviewing (CATI). The reference survey approach has been applied by several market research organizations, see for example Börsch-Supan *et al.* (2004) and Duffy *et al.* (2005).

Under the assumption of no non-response, or ignorable non-response, this reference survey will produce unbiased estimates of quantities that have also been measured in the web survey. Unbiased estimates for the target variable can be computed, but due to the small sample size,

these estimates will have a substantial variance. The question is now whether estimates can be improved by combining the large sample size of the web surveys with the unbiasedness of the reference survey.

To explore this, it is assumed that one qualitative auxiliary variable X is observed both in the web survey and in the reference survey, and that this variable has a strong correlation with the target variable Y of the survey. Then a form of post-stratification can be applied where the stratum means are estimated using web survey data and the stratum weights are estimated using the reference survey data. This leads to the post-stratification estimator

$$\bar{y}_{I,RS} = \sum_{h=1}^{L} \frac{m_h}{m} \bar{y}_I^{(h)}, \tag{25}$$

where $\bar{y}_I^{(h)}$ is the web survey based estimate for the mean of stratum h of the Internet population (for $h = 1, 2, \ldots, L$) and m_h/m is the relative sample size in stratum h for the reference sample (for $h = 1, 2, \ldots, L$). Under the conditions described above the quantity m_h/m is an unbiased estimate of $W_h = N_h/N$.

Let I denote the probability distribution for the web survey and let P be the probability distribution for the reference survey. Then the expected value of the post-stratification estimator is equal to

$$E(\bar{y}_{I,RS}) = E_I E_P(\bar{y}_{I,RS} \mid I) = E_I \left(\sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_I^{(h)} \right) = \sum_{h=1}^{L} W_h \bar{Y}_I^{(h)} = \tilde{Y}_I, \tag{26}$$

where $W_h = N_h/N$ is the relative size of stratum h in the target population. So, the expected value of this estimator is identical to that of the post-stratification estimator (19). The bias of this estimator is equal to

$$\begin{aligned} B(\bar{y}_{I,RS}) &= {}_I E(\bar{y}_{I,RS}) - \bar{Y} = \tilde{Y}_I - \bar{Y} = \sum_{h=1}^{L} W_h \left(\bar{Y}_I^{(h)} - \bar{Y}^{(h)} \right) \\ &= \sum_{h=1}^{L} W_h \frac{N_{NI,h}}{N_h} \left(\bar{Y}_I^{(h)} - \bar{Y}_{NI}^{(h)} \right). \end{aligned} \tag{27}$$

A strong relationship between the target variable and the auxiliary variable used for computing the weights means that there is little or no variation of the target variable within the strata. This implies that, if the stratum means for the Internet population and for the target population do not differ much, this results in a small bias. So, using a reference survey with the proper auxiliary variables can substantially reduce the bias of web survey estimates.

Note that the expression for the bias of the reference survey estimator is equal to that of the post-stratification estimator. An interesting aspect of the reference survey approach is that any variable can be used for adjustment weighting as long as it is measured in both surveys. For example, some market research organizations use "webographics" or "psychographic" variables that divide the population in "mentality groups." See Schonlau *et al.* (2004) for more details about the use of such variables.

People in the same groups have more or less the same level of motivation and interest to participate in such surveys. Effective weighting variables approach the MAR situation as much as possible. This implies that within weighting strata there is no relationship between participating in a web survey and the target variables of the survey.

Bethlehem (2007) shows that, if a reference survey is used, the variance of the post-stratification estimator is equal to

$$V(\bar{y}_{I,RS}) = \frac{1}{m}\sum_{h=1}^{L} W_h \left(\bar{Y}_I^{(h)} - \tilde{Y}_I\right)^2 + \frac{1}{m}\sum_{h=1}^{L} W_h(1-W_h)V\left(\bar{y}_I^{(h)}\right) + \sum_{h=1}^{L} W_h^2 V\left(\bar{y}_I^{(h)}\right). \quad (28)$$

The quantity $\bar{y}_I^{(h)}$ is measured in the web survey. Therefore its variance $V(\bar{y}_I^{(h)})$ will be of the order $1/n$. This means that the first term in the variance of the post-stratification estimator will be of the order $1/m$, the second term of order $1/mn$, and the third term of order $1/n$. Since n will generally be much larger than m in practical situations, the first term in the variance will dominate, that is the (small) size of the reference survey will determine the accuracy of the estimates. So, the large number of observations in the web survey does not help to produce precise estimates. One could say that the reference survey approach reduces the bias of estimates at the cost of a higher variance. See Section 4 for an example showing this effect.

3.4 Sampling the Non-Internet Population

If the target population of a survey is wider than just the Internet population, persons without Internet are excluded from a web survey. This problem could be solved by selecting a stratified sample. The target population is assumed to consist of two strata: the Internet population U_I of size N_I and the non-Internet population U_{NI} of size N_{NI}.

To be able to compute an unbiased estimate, a simple random sample must be selected from both strata. The web survey provides the data about the Internet-stratum. If this is a random sample with equal probabilities, the sample mean

$$\bar{y}_I = \frac{1}{n}\sum_{k=1}^{N} s_k I_k Y_k \quad (29)$$

is an unbiased estimator of the mean of the Internet population.

Now suppose a random sample (with equal probabilities) of size m is selected from the non-Internet stratum. Of course, there is no sampling frame for this population. This problem can be avoided by selecting a sample from the complete target population (a reference survey) and only using people without Internet access. Selected people with Internet access can be added to the large web sample, but this will have no substantial effect on estimators. The sample mean of the non-Internet sample is denoted by

$$\bar{y}_{NI} = \frac{1}{m}\sum_{k=1}^{N} t_k(1-I_k)Y_k, \quad (30)$$

where the indicator t_k denotes whether or not element k is selected in the reference survey, and

$$m = \sum_{k=1}^{N} t_k(1-I_k). \quad (31)$$

The stratification estimator is now defined by

$$\bar{y}_{ST} = \frac{N_I}{N}\bar{y}_I + \frac{N_{NI}}{N}\bar{y}_{NI}. \quad (32)$$

This is an unbiased estimator for the mean of the target population. Application of this estimator assumes the size N_I of the Internet population and the size N_{NI} of the non-Internet

population to be known. The variance of the estimator is equal to

$$V(\bar{y}_{ST}) = \left(\frac{N_I}{N}\right)^2 V(\bar{y}_I) + \left(\frac{N_{NI}}{N}\right)^2 V(\bar{y}_{NI}). \tag{33}$$

The variance of the sample mean in the Internet stratum is of order $1/n$ and the variance in the non-Internet stratum is of order $1/m$. Since m will be much smaller than n in practical situation, and the relative sizes of the Internet population and the non-Internet population do not differ that much, the second term will determine the magnitude of the variance. So the advantages of the large sample size of the web survey are for a great part lost by the bias correction.

The non-Internet population can also be sampled using a different approach, and that is a mixed-mode survey. A random sample is selected from the complete population. All selected people are given the choice to complete the questionnaire form on the Internet (if they have access) or to complete a paper form (if they do not have access).

3.5 Propensity Weighting

Propensity weighting is used by several market research organizations to correct for a possible bias in their web surveys. Examples can be found in Börsch-Supan *et al.* (2004) and Duffy *et al.* (2005). The original idea behind propensity weighting goes back to Rosenbaum & Rubin (1983, 1984). They developed a technique for comparing two populations. They attempt to make the two populations comparable by simultaneously controlling for all variables that were thought to explain the differences. In the case of a web survey, there are also two populations: those who participate in the web survey (if asked), and those who will not participate.

Propensity scores are obtained by modelling a variable that indicates whether or not someone participates in the survey. Usually a logistic regression model is used where the indicator variable is the dependent variable and attitudinal variables are the explanatory variables. These attitudinal variables are assumed to explain why someone participates or not. Fitting the logistic regression model comes down estimating the probability (propensity score) of participating, given the values of the explanatory variables.

Application of propensity weighting assumes some kind of random process determining whether or not someone participates in the web survey. Each element k in the population has a certain, unknown probability ρ_k of participating, for $k = 1, 2, \ldots, N$. Let $r_1, r_2, \ldots, r_N$ denote indicator variables, where $r_k = 1$ if person k participates in the survey, and $r_k = 0$ otherwise. Consequently, $P(r_k = 1) = \rho_k$.

The propensity score $\rho(X)$ is the conditional probability that a person with observed characteristics X participates, that is

$$\rho(X) = P(r = 1 \mid X). \tag{34}$$

It is assumed that within the strata defined by the values of the observed characteristics X, all persons have the same participation propensity. This is the MAR assumption that was introduced in Section 3.2. The propensity score is often modelled using a logit model:

$$\log\left(\frac{\rho(X_k)}{1-\rho(X_k)}\right) = \alpha + \beta' X_k + \varepsilon_k. \tag{35}$$

The model is fitted using Maximum Likelihood estimation. Once propensity scores have been estimated, they are used to stratify the population. Each stratum consists of elements with (approximately) the same propensity scores. If indeed all elements within a stratum have the same response propensity, there will be no bias if just the elements in the Internet population are used for estimation purposes. Cochran (1968) claims that five strata are usually sufficient

to remove a large part of the bias. Terhanian *et al.* (2001) describe one of the first applications propensity score weighting.

To be able to apply propensity score weighting, two conditions have to be fulfilled. The first condition is that proper auxiliary variables are available. These are variables that are capable of explaining whether or not someone is willing to participate in the web survey. Variables often used measure general attitudes and behaviour. They are sometimes referred to as "webographic" or "psychographic" variables. Schonlau *et al.* (2004) mention as examples "Do you often feel alone?" and "On how many separate occasions did you watch news programs on TV during the past 30 days?"

The second condition for this type of adjustment weighting is that the population distribution of the webographic variables must be available. This is generally not the case. A possible solution to this problem is to carry out an additional reference survey (see also Section 3.3). To allow for unbiased estimation of the population distribution, the reference survey must be 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 webographic 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. This is similar to the situation described in Section 3.3.

Schonlau *et al.* (2004) describe the reference survey of Harris Interactive. This is a CATI survey, using random digit dialling. This reference survey is used to adjust several web surveys. Schonlau *et al.* (2003) stress that the success of this approach depends on two assumptions: (1) the webographics variables are capable of explaining the difference between the web survey respondents and the other persons in the target population, and (2) the reference survey does not suffer from non-ignorable nonresponse. In practical situations it will not be easy to satisfy these conditions.

It should be noted that from a theoretical point of view propensity weighting should be sufficient to remove the bias. However, in practice the propensity score variable will often be combined with other (demographic) variables in a more extended weighting procedure. See Schonlau *et al.* (2004) for an example.

4 A Simulation Study

4.1 A Fictitious Population

To explore how effective correction techniques can be, a simulation study was carried out. A fictitious population was constructed. For this population, voting behaviour in the general elections was simulated and analysed.

To show more clearly what can happen, the relationships between the variables were taken somewhat stronger than in a real life situation. The literature on missing data (see for example Little & Rubin, 2002) distinguishes three mechanisms that cause data to be missing:

- *Missing completely at random* (MCAR). Missing data are caused by a variable Z that is completely unrelated to the target variable Y and any other variable X in the survey. Nonresponse is harmless. It only reduces the sample size. Estimates for parameters involving Y will not be biased.
- *Missing at random* (MAR). Missing data are partly by an auxiliary variable X. If there is a relationship between this variable X and the target variable Y, estimates for Y will be biased. Fortunately, it is possible to correct for such a bias by using a technique (for example weighting

adjustment) that takes advantage of the availability of the distribution of X in the sample or the population.
- *Not missing at random* (NMAR). Missing data are caused by a variable Z that is directly related to the target variable Y of the survey. This relationship cannot be accounted for by any auxiliary variable X. Estimates for Y will be biased. Unfortunately, correction techniques using X will not be able to remove the bias.

Two of these mechanisms are used in the simulation study. With respect to the Internet population, both MAR and not missing at random (NMAR) were introduced. The characteristics of estimators (before and after correction) were computed based on a large number of simulations.

First, the distribution of the estimator was determined in the ideal situation of a simple random sample from the target population. Then, it was explored how the characteristics of the estimator change if a simple random sample was selected just from the Internet population. Finally, the effects of weighting (post-stratification and reference survey) were analysed.

A fictitious population of 30,000 individuals was constructed. There were five variables:

- Age in three categories: young (with probability 0.40), middle aged (with probability 0.35) and elderly (with probability 0.25).
- Ethnic origin in two categories: native (with probability 0.85) and non-native (with probability 0.15).
- Having access to Internet with two categories (yes and no). The probability of having access to Internet depended on the two variables Age and Ethnic origin. For natives, the probabilities were 0.90 (for young), 0.70 (for middle aged) and 0.50 (for old). So, Internet access decreases with age. For non-natives, these probabilities were 0.20 (for young), 0.10 (for middle aged), and 0.00 (for old). These probabilities were chosen to reflect a much lower Internet access among non-natives.
- Voted for the National Elderly Party (NEP). The probability to do so depended on age. Probabilities were 0.00 (for young), 0.40 (for middle aged), and 0.60 (for elderly).
- Voted for the New Internet Party (NIP). The probability to do so depended on both age and having Internet. For people with Internet the probabilities were 0.80 (for young), 0.40 (for middle aged), and 0.20 (for elderly). For people without Internet all probabilities were equal to 0.10. For people with Internet voting for the NIP decreased with age. Only a few people without Internet voted for the NIP.

Figure 5 shows the relationships between various variables in a graphical way. The variable NEP suffers from Missing Completely At Random (MAR). There is direct relationship between voting behaviour and age, and also there is a direct relationship between age and having Internet. This will cause estimates to be biased. It should be possible, however, to correct for this bias by weighting using the variable age.

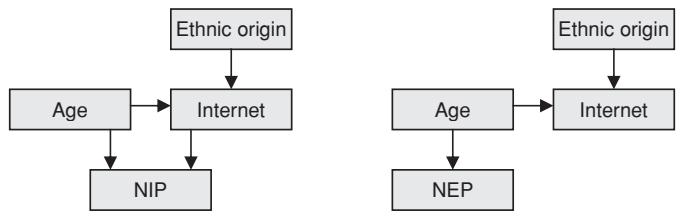

Figure 5. *Relationships between variables.*

The variable NIP suffers from NMAR. There exists (among other relationships) a direct relationship between voting for the NIP and having Internet. Estimates will be biased, and there is no correction possible.

4.2 Simulation Results for NEP

The distribution of estimators for the percentage of reported voters for both parties was determined in various situations by repeating the selection of the sample 1,000 times. In all cases the sample size was $n = 2{,}000$.

Figure 6 contains the results for the variable NEP. The left-most box plot shows the distribution of the estimator for simple random samples from the complete target population. The population value to be estimated is 25.4%. The estimator has a symmetric distribution around this value. This is a clear indication that the estimator is unbiased.

The second box plot shows what happens if samples are not selected from the complete target population, but just from the Internet population. The shape of the distribution remains the same, but the distribution as a whole has shifted downwards. All values of the estimator are systematically too low. The expected value of the estimator is only 20.3%. The 95% confidence interval lies, on average, between 18.6% and 20.3%. This interval does not contain the true value of 25.4%. So the confidence level is 0% instead of 95%. The estimator is clearly biased. The explanation of this bias is simple: relative few elderly have Internet. Therefore, they are under-represented in samples selected from the Internet. These are typically people who will vote for the NEP.

The third box plot in Figure 6 shows the distribution of the estimator in case of post-stratification by age. The bias is removed. This was possible because this is a case of MAR.

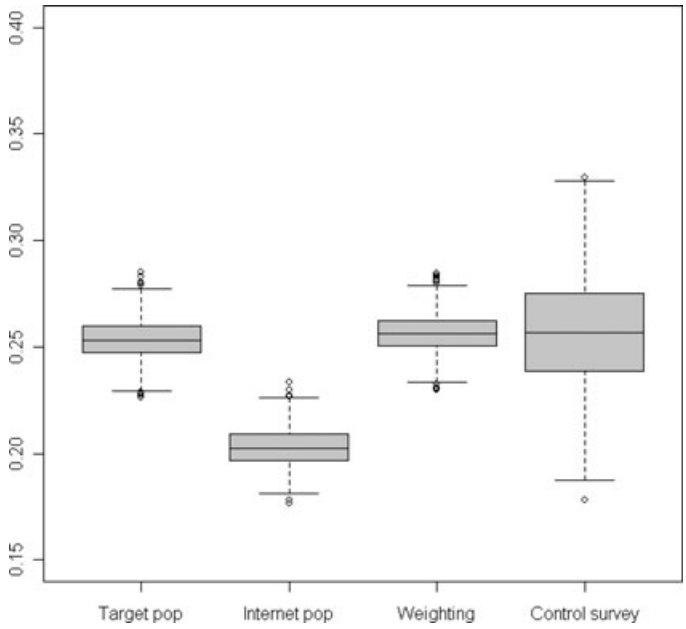

Figure 6. *Results of the simulations for variable NEP.*

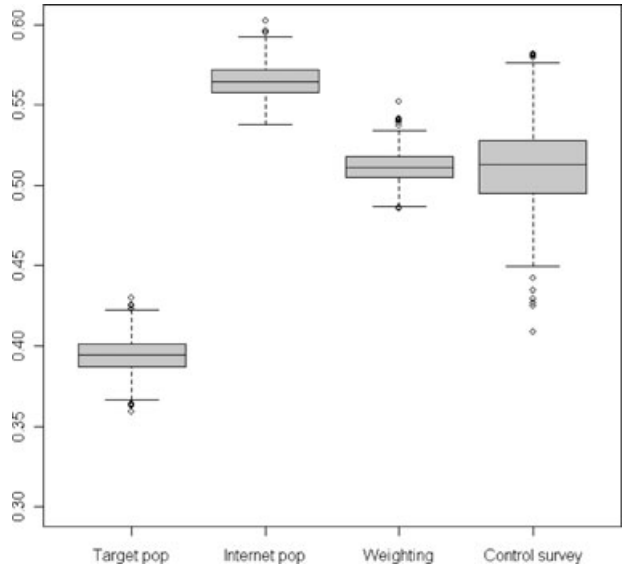

Figure 7. *Results of the simulations for variable NIP.*

Post-stratification by age can only be applied if the distribution of age in the population is known. If this is not the case, one could consider to conduct a small ($m = 100$) reference survey, in which this population distribution is estimated unbiasedly. The box plot on the right in Figure 6 shows what happens in this case. The bias is removed but at the cost of a substantial increase in variance.

4.3 Simulation Results for NIP

Figure 7 shows the results for the variable NIP. The box plot on the left shows the distribution of the estimator for simple random samples from the complete target population. The population value to be estimated is 39.5%. Since the estimator has a symmetric distribution around this value, it is clear that it is unbiased.

The second box plot shows what happens if samples are not selected from the complete target population, but just from the Internet population. The distribution has shifted upwards considerably. All values of the estimator are systematically too high. The expected value of the estimator is now 56.5%. The estimator is severely biased. The explanation of this bias is straightforward: NIP voters are over-represented among Internet users.

The third box plot in Figure 7 shows the effect of post-stratification by age. Only part of the bias is removed. This is not surprising as there is a direct relationship between reported voting for the NIP and having Internet. This is a case of NMAR.

Also in this case one can consider conducting a small reference survey if the population distribution of age is not available. The box plot on the right in Figure 7 shows what happens in this case. Again, only a part of the bias is removed and at the same time there is a substantial increase in variance.

4.4 Conclusions from the Simulation Study

The following conclusion can be drawn from this simulation study:

- If MAR or NMAR applies to Internet access, estimates based on a web survey will be biased.
- There is no guarantee that weighting will remove the bias. This correction technique will only work in case of MAR, and the proper auxiliary variables are used for weighting.
- A reference survey will only be effective in removing the bias if MAR applies, and the proper auxiliary variable measured.
- Use of a small reference survey will always substantially increase the variance of estimators. Or to say it differently: it will decrease the effective sample size considerably.

Bethlehem (2009a) describes a similar simulation experiment to explore the effects of self-selection. Response probabilities were modelled to depend on the same auxiliary variables as in this section. The conclusions were similar. Weighting will only remove the bias if MAR applies and the proper auxiliary variables are used for weighting. Using a small reference survey will substantially reduce the effective sample size in this situation

5 Using the Web for General Population Surveys

5.1 Single-Mode Web Surveys

Can the web be used as a mode of data collection in general population surveys, were focus is on obtaining precise and unbiased estimates of population characteristics? It has been argued in this paper that application of the principles of probability sampling is of crucial importance. This rules out self-selection surveys.

It is possible to conduct a web survey that is based on probability sampling. This requires a sampling frame. Sometimes such sampling frames are available. Examples are a survey among students of a university or employees of a company. The situation is not so straightforward for a general population survey. Unfortunately, there are no population registers containing e-mail addresses. A solution can be to approach sampled persons by some other mode. One option is to send them a letter with the request to go to a specific website, where they can complete the online questionnaire form. Such a letter should also contain a unique identification code that has to be entered. Use of such identifying codes guarantees that only sampled persons respond, and that they respond only once. Another option is to approach sampled persons face-to-face (CAPI) or by telephone (CATI) and asking them for their e-mail address. If they have, they are sent a link to the online questionnaire form.

There are people without access to Internet. Internet access by households varies between 19% and 83% in the European Union. So there is an under-coverage problem. Since there are differences between those with and without Internet access, under-coverage will often cause estimates to be biased. Internet penetration will increase over time. This helps to reduce the bias. However, it is not impossible that those without Internet will diverge (on average) more and more from those having Internet. Hence, there is no guarantee that problems will vanish in the near future.

Is it possible to conduct a web survey that is not affected by under-coverage and self-selection? The Dutch LISS panel is the result of such an attempt, see Scherpenzeel (2008). This online panel has been constructed by selecting a random sample of households from the population register of The Netherlands. Selected households were recruited for this panel by means of CAPI or CATI. So sample selection was based on true probability sampling. Moreover, co-operative households

without Internet access were provided with equipment giving them access to Internet. Analysis by Scherpenzeel & Bethlehem (2010) shows that the results of this panel are much closer to those based surveys based on probability sampling than to those of surveys using self-selection web surveys.

A possible lest costly solution for those without Internet access at home is to ask them to go to an Internet café, public library, or to complete the survey at their workplace. A problem that may remain is that some will not have the skills or experience to work with computers. Possible problems may be blindness, lack of proper colour perception, cognitive limitations, and illnesses like Parkinson's disease.

It should be noted that also other modes of data collection have their coverage problems. For example, a CATI survey requires a sampling frame consisting of telephone numbers. Statistics Netherlands can use only listed telephone numbers for this. Almost all of these numbers are fixed-line numbers. Only between 60% and 70% of the people in the Netherlands have a listed phone number, see Cobben (2004).

The under-coverage problem for CATI surveys will become even more severe over time. This is due to the popularity of mobile phones and the lack of lists of mobile phone numbers, see for example Kuusela (2003). The situation is improving for web surveys. In many countries there is a rapid rise in households having Internet access. For example, the percentage of households with Internet is now over 80% in The Netherlands, and it keeps growing. So one might expect that in the near future web survey coverage problems will be less severe.

5.2 Mixed-Mode Web Surveys

Budget cuts on the one hand and demands for more and more detailed information, while maintaining an acceptable level of data quality, have stimulated statistical agencies in several countries to explore different approaches to data collection. One such approach is the mixed-mode survey. Different data collection modes are used in such a survey. De Leeuw (2005) describes two-mixed approaches. The first approach is use of different mode *concurrently*. The sample is divided into groups and each group is approached by a different mode. The other approach is use of different modes *sequentially*. All sample persons are approached by one mode. The non-respondents are then followed up by a different mode than the one used in the first approach. This process can be repeated for a number of modes.

If cost reduction is the main issue, one could think of a mixed-mode survey that starts with a questionnaire on the web. Non-respondents are followed up by CATI. Non-respondents remaining after CATI could be followed up by CAPI. So the survey starts with the cheapest mode and ends with the most expensive one.

If quality and response rate are of vital importance, one could think of a mixed-mode design that starts with CAPI. The non-response is followed-up by CATI. Remaining non-respondents are asked to complete the questionnaire on the web.

Mixed-mode surveys suffer from mode effects. This is the phenomenon that the same question is answered differently when asked in a different mode. One can attempt to design the survey in such a way that mode effects are minimized as much as possible. Dillman *et al.* (2008) propose the *unimode* approach. It is a set of questionnaire design guidelines. Application of these guidelines should lead to questions that are understood in the same way and are answered in the same way in different modes.

Statistics Netherlands is considering implementing mixed-mode designs for some of its surveys. These surveys will start with the web-mode and subsequent modes will be CATI and CAPI. Note that changing to a mixed-mode survey will also have consequences for repeated surveys, as it may lead to breaks in time-series.

5.3 Conclusions

Web surveys are an interesting new way of survey data collection. At first sight, a web survey seems to have attractive properties. It provides simple, cheap and fast access to a large group of potential respondents. There are, however, also methodological issues, like under-coverage and self-selection. This raises the question whether web surveys can be used for data collection in official statistics.

If the objective is to obtain accurate estimates of population characteristics, application of the probability sampling paradigm is vital. Consequently, there is no role for self-selection web surveys. It has been shown that such surveys may result in severely biased estimates.

Currently, under-coverage also still is a serious problem a many countries, as Internet penetration is not high everywhere. Research shows there can be substantial differences between those having access to Internet and those having no access, which also results in biased estimates.

It was explored whether weighting adjustment techniques may be able to reduce a bias to the under-coverage or self-selection. Like in non-response correction techniques, this only works if the proper auxiliary variables are available. This is often not the case. In interesting possibility is to conduct a small reference survey to collect those variables. This only works if a different data collection mode (e.g. CAPI or CATI) is used and there is no non-response or only ignorable non-response. Such a reference survey may be capable of reducing the bias, but a price has to be paid: the standard errors of the estimates will dramatically increase if the sample size of the reference survey is mall. Consequently, the advantages of the large sample size of the web survey are lost.

References

Bethlehem, J.G. (1988). Reduction of the nonresponse bias through regression estimation. *J. Official Statist.*, **4**, 251–260.

Bethlehem, J.G. (2002). Weighting nonresponse adjustments based on auxiliary information. In *Survey Nonresponse*, Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge & R.J.A. Little. New York: Wiley & Sons.

Bethlehem, J.G. (2007). Reducing the bias of web survey based estimates. Discussion paper 07001. Voorburg/Heerleen, The Netherlands: Statistics Netherlands.

Bethlehem, J.G. (2009a). *Applied Survey Methods, A Statistical Perspective*. Hoboken, NJ: John Wiley & Sons.

Bethlehem, J.G. (2009b). *The Rise of Survey Sampling*. Discussion Paper 09015. The Hague/Heerlen, The Netherlands: Statistics Netherlands.

Börsch-Supan, A., Elsner, D., Faßbender, H., Kiefer, R., McFadden, D. & Winter, J. (2004). *Correcting the Participation Bias in an Online Survey*. Report. Munich, Germany: University of Munich.

Bowley, A.L. (1906). Address to the economic science and statistics section of the British association for the advancement of science. *J. R. Stat. Soc.*, **69**, 548–557.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bull. Int. Statist. Inst.*, **XII**, Book 1, 6–62.

CBS (2008). ICT gebruik van huishoudens naar huishoudkenmerken. Statline. www.cbs.nl

Cobben, F. (2004). *Nonresponse Correction Techniques in Household Surveys at Statistics Netherlands: a CAPI-CATI Comparison*. Technical report, Voorburg, The Netherlands: Statistical Netherlands, Methods and Informatics Department.

Cochran, W.G. (1953) *Sampling Techniques*. New York: John Wiley & Sons.

Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 205–213.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, **64**, 464–494.

Couper, M.P. (2008). *Designing Effective Web Surveys*. Cambridge, UK: Cambridge University Press.

Couper, M.P., Baker, R.P., Bethlehem, J.G., Clark, C.Z.F., Martin, J., Nicholls II, W.L. & O'Reilly, J.M. (Eds.) (1998). *Computer Assisted Survey Information Collection*. New York: Wiley & Sons.

De Haan, J. & Van 't Hof, C. (2006). *Jaarboek ICT en Samenleving, De Digitale Generatie*. The Hague, The Netherlands: Netherlands Institute for Social Research/SCP.

De Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *J. Official Statist.*, **21**, 233–255.
Deming, W. E. (1950). *Some Theory of Sampling*. New York: John Wiley & Sons.
Dillman, D. A. & Bowker, D. (2001). The web questionnaire challenge to survey methodologists. In *Dimensions of Internet Science*, Eds. U.D. Reips & M. Bosnjak. Lengerich, Germany: Pabst Science Publishers, 159–178.
Dillman, D.A., Smyth, J.D. & Christian, L.M. (2008). *Internet, Mail, and Mixed-Mode Surveys, The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons.
Duffy, B., Smith, K., Terhanian, G. & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *Internat. J. Market Res.*, **47**, 615–639.
Eurostat (2007). *More than 40% of Households have Broadband Internet Access*. Eurostat News Release 166/2007. Luxembourg: Eurostat.
Fricker, R. & Schonlau, M. (2002). Advantages and disadvantages of internet research surveys: Evidence from the literature. *Field Meth.*, **15**, 347–367.
Hansen, M.H., Hurvitz, W.N. & Madow, W.G. (1953). *Survey Sampling Methods and Theory*. New York: John Wiley & Sons.
Heerwegh, D. & Loosveldt, G. (2002). *An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys*. In *Proceedings of the Paper Presented at the International Conference on Improving Surveys*, Copenhagen.
Horvitz, D.G. & D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685.
Kendall, M.G. (1960). Where shall the history of statistics begin? *Biometrika*, **47**, 447–449.
Kiaer, A. N. (1895), Observations et Expériences Concernant des Dénombrements Représentatives. *Bull. Internat. Statist. Inst.*, **IX**, Book 2, 176–183.
Kiaer, A. N. (1997 reprint). Den Repräsentative Undersökelsesmetode. *Christiania Videnskabsselskabets Skrifter. II. Historiskfilosofiske klasse*, **4**, 37–56.
Kish, L. (1967). *Survey Sampling*. New York: Wiley & Sons.
Kruskal, W. & Mosteller, F. (1979a). Representative sampling, I: Non-scientific literature. *Internat. Statist. Rev*, **47**, 13–24.
Kruskal, W. & Mosteller, F. (1979b). Representative sampling, II: Scientific literature. Excluding statistics. *Internat. Statist. Rev.*, **47**, 111–127.
Kruskal, W. & Mosteller, F. (1979c). Representative sampling, III: The current statistical literature. *Internat. Statist. Rev.*, **47**, 245–265.
Kuusela, V. (2003). Mobile phones and telephone survey methods. In *Proceedings of the 4th ASC International Conference*, Eds. R. Banks, J. Currall, J. Francis, L. Gerrard, R. Kahn, T. Macer, M. Rigg, E. Ross, S. Taylor & A. Westlake, pp. 317–327. Chesham Bucks, UK: Association for Survey Computing (ASC).
Little, R.J.A & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons.
Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc.*, **97**, 558–625.
Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rosenbaum, P.R. & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.*, **79**, 516–524.
Särndal, C.E. & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester, UK: John Wiley & Sons.
Scherpenzeel, A. (2008). An online panel as a platform for multi-disciplinary research. In *Access Panels and Online Research, Panacea or Pitfall?* Eds. I. Stoop & M. Wittenberg, pp. 101–106. Amsterdam: Aksant.
Scherpenzeel, A. & Bethlehem, J. (2010). How representative are online-panels? Problems of coverage and selection and possible solutions. In *Social Research and the Internet: Advances in applied Methods and New Research Strategies*. Eds. M. Das, P. Ester & L. Kaczmirek. New York: Routlegde Academic, in press.
Schonlau, M., Fricker, R.D. & Elliott, M.N. (2003). *Conducting Research Surveys via E-mail and the Web*. Santa Monica, CA: Rand Corporation.
Schonlau, M., Zapert, K., Payne Simon, L., Haynes Sanstad, K., Marcus, S., Adams, J. Kan, H., Turber, R. & Berry, S. (2004). A comparison between responses from propensity-weighted Web survey and an identical RDD survey. *Social Science Comp. Rev.*, **22**, 128–138.
Terhanian, G., Smith, R., Bremer, J. & Thomas, R.K. (2001). Exploiting analytical advances: Minimizing the biases associated with Internet-based surveys of non-random samples. *ESOMAR Publication Services*, **248**, 247–272.
Vonk, T., Van Ossenbruggen, R. & Willems, P. (2006). The effects of panel recruitment and management on research results, a study among 19 online panels. *ESOMAR Publication Services*, **317**, 79–99.
Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Charles Griffin & Co.

Résumé

A première vue, les enquêtes en ligne semble être un moyen intéressant et attrayant de collecte de données. Ils fournissent un accès simple, rapide et peu coûteux à un grand groupe de répondants potentiels. Les enquêtes en ligne ne sont toutefois pas sans problèmes méthodologiques. Certains groupes au sein des populations sont sous-représentés parce qu'ils ont un accès restreint à Internet. En outre, le recrutement des personnes interrogées est souvent fondé sur l'auto-sélection. Sous-couverture et auto-sélection peuvent toutes deux être la source d'estimations biaisées. Le présent article décrit ces problèmes méthodologiques. Il explore également l'effet des différentes techniques de correction (pondération corrective et utilisation d'enquêtes de référence). Tout ceci amène à la question de savoir si des enquêtes en ligne bien conçues sont propres à être utilisées pour la collecte de données. L'article tente de répondre à cette question. Il conclut que le problème de la sous-couverture pourrait se résoudre dans l'avenir, mais que l'auto-sélection aboutit à des résultats d'enquête non fiables.

[*Received September 2007, accepted February 2010*]