

Online Panels – A Paradigm Theft?

Jelke Bethlehem & Ineke Stoop

Abstract

Decreasing survey participation leads to increasing survey costs and smaller precision of survey estimates. In the likely case that noncontacts and refusals differ from respondents, it may also increase nonresponse bias. Online panels are promoted as the solution to these problems. Online panels are relatively inexpensive as large samples can easily be drawn from panels containing hundreds of thousands of willing respondents. With an increasing internet penetration, even specific groups such as the elderly will have internet access and therefore they can be included in online panels.

Survey methodology has been developed over a period of more than 100 years. The paradigm of probability sampling has shown to work well in social research, official statistics and market research. It has allowed researchers to produce well-founded and reliable survey results. Often the impression is created that this paradigm also applies to online panel research. Sadly, general survey quality criteria such as sample size and response rate cannot be generalized to panel studies. Under-coverage and self-selection may seriously limit their value for scientific and policy making purposes.

Some survey researchers claim that the problems mentioned above can be reduced by applying some kind of weighting adjustment procedure, e.g. using weighting variables measured in a reference survey. We argue this is a too optimistic point of view. The presentation will outline why online panels do not solve the problems caused by decreased participation. Will discuss online panels within a quality framework and will present examples of possibly useful applications of online panels in academic and governmental research.

In market research online panels are considered to be the future. There are signs, however, that despite the rapidly increasing market share of web research, the rising star of online panels is already on the wane.

Keywords

Access panels, online research, coverage, probability sampling, nonresponse.

1. The changing landscape of survey research

The survey research landscape has undergone radical changes over the last decades. First, there was the change from traditional paper and pencil interviewing to computer-assisted

interviewing. And now face-to-face, mail and telephone surveys are increasingly replaced by online surveys.

The popularity of online research is not surprising. An online survey is a simple means to get access to a large group of people. 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. And online surveys offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation and movies).

However, there is another side to this coin. Online research is not without methodological problems. These problems have an impact on the quality of the survey results. The cause of these problems can partly be found in the incorrect application of the principles of sample survey theory that have been developed more than a century ago.

The theory of survey sampling is heavily based on the probability sampling paradigm. By selecting random samples probability theory can be applied, making it possible to quantify accuracy of estimates. This paradigm has been successfully applied in official and academic statistics in the 1940's, and to a much lesser extent also in more commercial market research.

At first sight, online 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. What is different for many online surveys, however, is that the principles of probability sampling have not been applied. This can have a major impact on survey results.

It is not always clear what is meant by online research and how online surveys fit in the framework of survey sampling theory. There is also confusion about what it means. Online research is often used as a synonym for online panels, but this is a too limited view. There are also large online cross-sectional surveys. An example is the *21minuten.nl*, a survey supposed to supply answers to questions about important problems in Dutch society. Within a period of six weeks in 2006 about 170.000 people completed the online questionnaires. A similar survey was conducted in Germany (*Perspektive Deutschland*).

Panels need not necessarily be online panels. Panels existed already long before the Internet emerged. Panel members can very well complete questionnaires face-to-face, by telephone or by mail. An interesting example of an online panel 'avant la lettre', is the Telepanel, see Saris (1998). It started in 1986 and used home computers placed in the homes of panel members. Questionnaires were downloaded, and answers uploaded, by means of telephone and modem.

Online research is often claimed to be representative because of the high number of respondents or as a result of advanced adjustment weighting procedures. The term representative is rather confusing. It can have many meanings and is often used in a very loose sense to convey a vague idea of good quality. A high number of respondents are often considered to ensure validity and reliability. There are serious doubts, however, whether a large sample size as a result of self-selection of respondents has the same meaning as a large sample size in probability sampling. Similarly, a high response rate in a sample among cooperative panel members is unlikely to have the same impact on the quality of outcomes as a high response rate in a random sample of the population.

This paper discusses online research, web surveys and access panels from the perspective of classic survey quality criteria and compares the different paradigms behind probability sampling and other forms of sample selection. It ends with a plea for transparency.

2. So you have a representative sample?

In her overview of the history of survey research in the United States, Converse (1987) sketches the permanent controversy between market research organizations with their tradition of non-probability samples such as quota sampling and ‘juries’, and official statistics with probability sampling founded in inferential statistics and probability theory. Fowler (2002, p. 53) describes the gap between both traditions as follows: ‘The federal government will not fund survey research efforts designed to make estimates of population characteristics that are not based on probability sampling techniques. Most academic survey organizations and many non-profit research organizations have a similar approach to sampling. At the same time, most of the major public opinion groups, political polling groups, and market research organizations rely heavily on non-probability sampling methods’.

The controversy goes back to the roots of survey sampling, see e.g. Kish (2003). Anders Kiaer, the director of the Norwegian Statistical Bureau, can be seen as the founder of the survey method that is now widely applied in official statistics and social research. In 1895 he published his *Representative Method*. It was a partial inquiry in which a large number of persons were questioned. This selection should form 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 *representativeness*. 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), who made the first steps in this direction. He showed that for large samples, selected at random from the population, the estimate 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 representativeness 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.

The contribution of Neyman was not only that he invented the confidence interval. By making an empirical evaluation of Italian census data, he could prove that the Representative Method based on purposive sampling failed to provide satisfactory estimates of population characteristics. The result of Neyman’s evaluation of purposive sampling was that the method fell into disrepute in official statistics.

The concept of ‘representativity’ plays a crucial role in the discussion about the foundations of survey sampling. Kruskal and Mosteller (1979a, 1979b and 1979c) present an extensive overview of what representative is supposed to mean in non-scientific literature, scientific literature excluding statistics and in the current statistical literature. They found the following meanings for ‘representative sampling’: (1) general acclaim for data, (2) absence of selective

forces, (3) miniature of the population, (4) typical or ideal case(s), (5) coverage of the population, (6) a vague term, to be made precise, (7) representative sampling as a specific sampling method, (8) as permitting good estimation, or (9) good enough for a particular purpose.

Kruskal and Mosteller (1979b, p. 125) recommended not using the word *representative*, but instead to specify what one means. This did not catch on. Kish (2003, p. 12) thought that *representative sampling* was a term that can be avoided and assumed in 1995 that it was disappearing from the technical vocabulary. Schnell (1997, p. 12) called it an immeasurable, unscientific concept, both with little success. ‘*Representative*’ in The Netherlands is still a household word in describing samples and sampling even when it is not clear whether this means a miniature of the population, or good coverage of the population, or a random sample or just “trust me!”.

The problem is, of course, that both in probability sampling and other forms of sampling claims are made samples are representative, often with quite different meanings and sometimes with no concrete meaning at all besides conveying a vague sense of good quality.

3. By a small sample we may judge the whole piece?

The basics of probability sampling as it is applied now in e.g. official statistics are laid down by Horvitz and Thompson (1952) in their seminal paper. They state that unbiased estimators of population characteristics can always be constructed, provided samples are selected by means of probability sampling and every element in the population has a known and strictly positive probability of being selected.

Moreover, under these conditions standard errors of estimates, and thus confidence intervals, can be computed. Therefore it is possible to establish the accuracy of estimates. The Horvitz-Thompson approach can be used also in surveys with complex sampling designs, like stratified random samples, cluster samples and two-stage samples.

Notwithstanding the advantages of probability sampling, there are many forms of non-probability sampling. Kalton (1983) distinguishes three types of non-probability samples:

- *Convenience samples*. Sample selection is mainly based on easy availability or accessibility of elements. An example is conducting survey interviews in a shopping mall on Saturday afternoon. Also samples composed of respondents volunteering to participate are convenience samples.
- *Purposive samples*. The sample is chosen by a subject matter expert in such a way that is ‘representative’ in his/her opinion. The expert will usually attempt to include elements that cover all various aspects in the population.
- *Quota sampling*. Interviewers are given quotas of different types of people with whom they have to conduct interviews. Quotas are often based on demographic characteristics like gender, age, marital status and neighbourhood. Quota sampling is similar to the original Representative Method developed by Anders Kiaer in 1895.

The weakness of all non-probability methods is that there is no theoretical framework although efforts have been made to develop this (Deville, 1991). Therefore it is not possible to establish accuracy of estimates other than by subjective assessment. Nevertheless, despite its theoretical weakness, non-probability sampling is widely used, particularly in market

research. Usually, the reasons to use it are convenience and relatively low costs. Increasingly, market researchers seem to feel that the problems facing regular surveys (declining participation, people being at home less often, legal restrictions, rapidly increasing costs) are such that non-probability sampling – especially in the form of online panels – is actually to be preferred above traditional surveys or at least considered as an equal and more manageable alternative (see for instance Thomas, 2006) in his editorial to the recent ESOMAR Panel Research Conference.

Morton-Williams (1993, pp.31-35) has shown that the claim that quota samples are representative is based on two assumptions, namely ‘... that the behaviour and attitudes to be measured are related primarily to the variables used as quota controls; secondly, that they are not associated independently of these controls with factors underlying nonresponse nor with the characteristics of those likely to require more than one call to obtain an interview’ (p.32). Or, as Sudman (1966) said a long time ago: ‘In probability sampling with quotas the basic assumption made is that it is possible to divide the respondents into strata in which the probability of being available for interviewing is known and is the same for all individuals within the stratum, although varying between strata.’ These assumptions cannot be put to the test. Similar claims, however, are made with respect to access panels.

Online panels, also called *access panels*, and for short just *panels*, are becoming increasingly popular in the western world. A panel consists of persons who have agreed to regularly participate in surveys run by a specific organization, generally a market research organization. There are various ways to set up an online panel. Both probability and non-probability sampling techniques can be used.

Ideally, a panel is constructed using a random sample from a population. One way to select such a sample is using Random Digit Dialling. Another is to invite respondents in a probability based survey to become a member of a follow-up panel. If the original selection probabilities and response rates are available (which is rarely the case), a survey among (a random sample from) an access panel can theoretically be considered as a case of probability sampling. In this case, an access panel is very similar to a host survey.

Unfortunately, many online panels are based on some form of non-probability sampling. Major opinion polls in The Netherlands rely on self-selection of respondents. The same is true for the large 21minutes.nl web survey. A study across 19 online panels of Dutch market research organisations shows that most of them use self-registration, links and banners on websites or snowballing, see Vonk et al. (2006). This all means that most online research has two fundamental methodological flaws:

Keywords *Under-coverage*. People without Internet will never participate in online research. This means research results can only apply to the Internet population and not to the complete population.

Keywords *Self-selection*. Researchers have no control over the selection mechanism. Selection probabilities are unknown. Therefore, no unbiased estimates can be computed, nor can the accuracy of estimates be established.

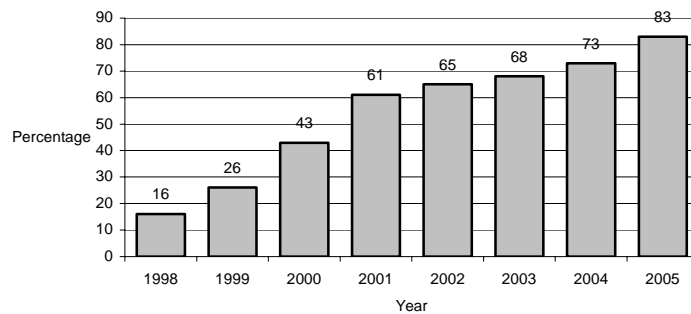
These two problems are analysed in more detail in the subsequent sections.

4. Coverage problems

Online research suffers from under-coverage because the target population is usually much wider than the Internet population. Bethlehem (2007) has analysed the situation in the Netherlands.

The percentage of persons having an Internet connection at home increases from year to year, see figure 4.1. In seven years time, the percentage of Internet connections increased from 16% to 83%. Still, it is clear that not every household will have access to Internet in the near future.

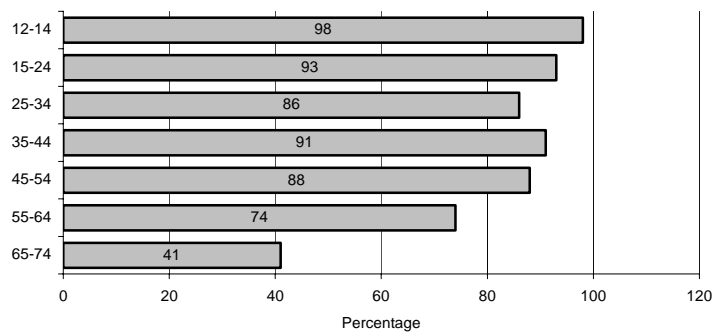
Figure 4.1. Percentage of persons having Internet



Internet access is unevenly distributed over the population. More males than females have access to the Internet. Figure 4.2 shows the percentage of people having Internet access at home by age group (in 2005). This percentage decreases with age. Particularly, the elderly are much under-represented when the Internet is used as a selection mechanism.

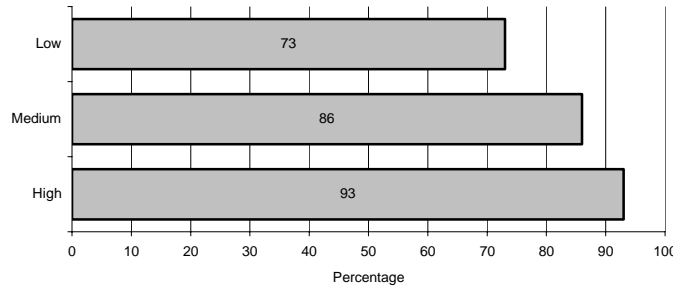
Figure 4.3 contains the percentage of people using the Internet by level of education (in 2005). It is clear that people with a higher level of education more frequently have Internet than people with a lower level of education.

Figure 4.2. Having Internet, by age.



Internet access among non-native young people is much lower 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. The results described above are in line with the findings of authors in other countries. See e.g. Couper (2000), and Dillman and Bowker (2001).

Figure 4.3. Having Internet by level of education.



To see what the impact of under-coverage on estimates can be, we analyse the situation in which a proper random sample is selected from the Internet population. Let the target population of the survey consist of N persons, which are labelled $1, 2, \dots, N$. Associated with each person k is a value Y_k of the target variable Y . Aim of the online survey is assumed to be estimation of the population mean

$$Y = \frac{1}{N} \sum_{k=1}^N Y_k \tag{4.1}$$

of the target variable Y .

The population U is divided into two sub-populations U_I of persons having access to Internet, and U_{NI} of persons not having access to the Internet. Associated with each person k is an indicator I_k , where $I_k = 1$ if person k has access to the Internet (and thus is a member of sub-population U_I), and $I_k = 0$ otherwise. The sub-population U_I will be called the *Internet population*. Let

$$N_I = \sum_{k=1}^N I_k \tag{4.2}$$

denote the size of sub-population U_I . Likewise, N_{NI} denotes the size of the sub-population U_{NI} , where $N_I + N_{NI} = N$. The mean of the target variable for the elements in the Internet-population is equal to

$$\bar{Y}_I = \frac{1}{N_I} \sum_{k=1}^N I_k Y_k . \tag{4.3}$$

A random sample selected without replacement from the Internet-population can be represented by a series a_1, a_2, \dots, a_N of N indicators, where the k -th indicator a_k assumes the value 1 if person k is selected, and otherwise it assumes the value 0, for $k = 1, 2, \dots, N$. Note that always $a_k = 0$ for elements k outside the Internet-population. The sample size is denoted by n_I . The sample mean

$$\bar{y}_I = \frac{1}{n_I} \sum_{k=1}^N a_k I_k Y_k \tag{4.4}$$

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

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

(4.5)

The magnitude of this bias is determined by two factors. The first factor is the relative size N_{NI} / N of the sub-population without Internet. The bias will increase as a larger proportion of the population does not have access to Internet. The second factor is 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.

Since Internet coverage is steadily increasing, the factor N_{NI} / N is decreasing. This has a bias reducing effect. However, it is not clear whether the contrast also decreases. To the contrary, it is not unlikely that the (small) group of people without Internet will be more and more different from the rest of the population. As a result, substantial bias may still remain.

5. Self-selection problems

The problem of self-selection is that the researcher has no control over the selection mechanism of the survey. Respondents are those people who happen to have Internet, visit the website and decide to participate in the survey. Therefore, no unbiased estimates can be computed nor can the accuracy of estimates be determined.

Table 5.1. Dutch Parliamentary elections 2003
Outcomes and the results of various opinion surveys

	Election	Kennisnet	RTL4	SBS6	Nederland 1
Sample size		17,000	10,000	3,000	1,200
Seats in parliament:					
CDA (christian democrats)	44	29	24	42	42
LPF (populist party)	8	18	12	6	7
VVD (liberals)	28	24	38	28	28
PvdA (social democrats)	42	13	41	45	43
SP (socialists)	9	22	10	11	9
GL (green party)	8	26	9	6	8
D66 (liberal democrats)	6	4	7	5	6
Other parties	5	14	9	7	7
Mean Absolute Difference		12.5	5.3	1.8	0.8

The effects of self-selection can be illustrated using an example related to the general elections in The Netherlands in 2003. Various organisations made attempts to use opinion polls to predict the outcome of these elections. The results of these polls are summarised in table 5.1.

A typical example of a self-selecting survey was the survey on the Dutch website *Kennisnet* (Knowledge net). This is a website for all those involved in education. More than 11,000 schools and other educational institutes use this website. The survey was an opinion poll for the general elections of 22 January 2003. Everybody, also those not involved in education, could participate in the poll. Table 4.1 contains both the official results (seats in parliament) of the election (column *Election*) and the results of this poll on the morning of the Election Day (column *Kennisnet*). The survey estimates were based on votes of approximately 17,000

people. No adjustment weighting was carried out. Although this is a large sample, it is clear that the survey results were no way near the true election results. The Mean Absolute Difference (MAD) is equal to 12.5, which means that the estimated number of seats and the true number of seats differ on average by an amount of 12.5. This survey could certainly not be used for predicting election results.

Another example of a self-selection web survey was the election site of the Dutch Television channel RTL 4. It resembled to some extent the Kennisnet survey, but was targeted at a much wider audience. Again, the survey researcher had no control at all over who was voting. There was some protection, by means of cookies, against voting more than once. However, this also had the draw-back, that only one member of the family could participate. Table 5.1 shows the survey results at noon on the day of the general elections (column *RTL4*). Figures were based on slightly over 10,000 votes. No weighting adjustment procedure was carried out. The results are better than that of the Kennisnet survey (the MAD decreased from 12.5 to 5.3). However, deviations between estimates and true figures are still substantial, particularly for the large parties. Note that even a large sample size of over 10,000 people does not help to get accurate estimates.

The Dutch commercial television channel *SBS6* used an access panel. Values of basic demographic variables were available for all panel members. A sample of size 3,000 was selected. Selection was carried out such that the sample was representative with respect to the social-demographic and voting characteristics. Table 5.1 shows the results (column *SBS6*). The survey took place on the day before the general elections. Although attempts have been made to create a 'representative' sample, the results differ still from the final result. The MAD has decreased to 1.8, but is still substantial.

A better prediction was obtained with a true probability sample. The table shows the results of a survey based on such a probability sample. It was carried out by the television channel *Nederland 1* in co-operation with the marketing agency *Interview-NSS*. A sample of size 1,200 was selected by means of random digit dialling. The MAD was reduced to 0.8.

Table 5.2. Dutch Parliamentary elections 2006
Outcomes and the results of various opinion surveys

	Election result	Politieke¹⁾ Barometer	Peil.nl²⁾	De Stemming³⁾	DPES 2006⁴⁾
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

1) Politieke Barometer, Interview-NSS: sample from online panel (N=1000)

- 2) Peil.nl, Maurice de Hond: representative sample from online panel (weighted) (N=2500)
- 3) De Stemming, TNS-NIPO: sample from online panel (N=2000)
- 4) Dutch Parliamentary Election Study, fieldwork 10 October-22 November 2006, two-stage random sample from population register, N=4000, response rate=65%.

A more recent comparison is presented in table 5.2. Sample sizes are similar in this case and the differences between MAD based on three samples from online panels and one random sample are much smaller. As in 2003 the random sample, in this case the Dutch Parliamentary Election Study conducted by Statistics Netherlands, clearly outperformed the other surveys.

The conclusion from the analysis above is that a probability sample is a vital prerequisite for making proper inference about the target population of a survey. Even with a probability sample of only size 1,200 better results can be obtained than with a non-probability sample of size 10,000 or more. To explore the effect of self-selection on estimates, we assume that each person k in the Internet-population has unknown probability ρ_k of participating in the survey, for $k = 1, 2, \dots, N_I$. The responding persons can be denoted by a series r_1, r_2, \dots, r_N of N indicators, where the k -th indicator r_k assumes the value 1 if person k participates, and otherwise it assumes the value 0. Note that sampling without replacement is assumed. The expected value $\rho_k = E(r_k)$ is called the *response propensity* of person k . For sake of convenience we have also introduced response propensities for non-Internet-population elements. By definition the values of all these probabilities are 0. The realised sample size is denoted by $n_S = r_1 + r_2 + \dots + r_N$.

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

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^N r_k Y_k \quad (5.1)$$

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

$$E(\bar{y}_S) \approx \bar{Y}_I^* = \frac{1}{N_I \bar{\rho}} \sum_{k=1}^N \rho_k I_k Y_k \quad (5.2)$$

where $\bar{\rho}$ is the mean of all response propensities in the Internet-population, see Bethlehem (2007).

Generally, the expected value of the sample mean is not equal to the population mean of the Internet-population. The only situation in which the bias vanishes is that in which all response propensities in the Internet-population are equal. Indeed, in this case, self-selection does not lead to an unrepresentative sample because all elements have the same selection probability.

Bethlehem (1988) shows that the bias of the sample mean (5.1) can be written as

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y}_I \approx \bar{Y}_I^* - \bar{Y}_I = \frac{C(\rho, Y)}{\bar{\rho}}, \quad (5.3)$$

in which

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

(5.4)

is the covariance between the values of target variable and the response propensities in the Internet-population. The bias of the sample mean (as an estimator of the mean of the Internet population) is determined by two factors:

- The average response propensity. The more likely people are to participate in the survey, the higher the average response propensity will be, and thus the smaller the bias will be.
- The relationship between the target variable and response behaviour. The stronger the relationship, the higher the bias will be.

Three situations can be distinguished in which this bias vanishes:

- 1) All response probabilities are equal. Again, this is the case in 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. If this were the case, 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.

If it is the objective of the survey to estimate the mean of the total population (and not just the mean of the Internet population), two factors contribute to the bias: under-coverage and self-selection. Although it is theoretically possible that these two effects compensate one another, it is more likely in many practical situations that they enforce each other.

6. Does weighting adjustment help?

Weighting adjustment is a family of techniques that attempt to improve the quality of survey estimates by making use 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 is available. By comparing the population distribution of an auxiliary variable with its sample 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 can now be obtained by using the weighted values instead of the unweighted values. Weighting adjustment is used to correct surveys that are affected by nonresponse, see e.g. Bethlehem (2002).

Post-stratification is a well-known and often used weighting method. To carry out post-stratification, one or more qualitative auxiliary variables are needed. Together they divide the target population into a number of strata (i.e. sub-populations). Identical adjustment weights are assigned to all elements in the same stratum. The bias of the estimate based on weighted data 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 and the stratification variables. Then the variation in the values of the target variable manifests itself between strata but not within strata. In other words, the strata are

homogeneous with respect to the target variable. Unfortunately, such auxiliary variables are not very often available, or there is only a weak correlation.

If proper auxiliary variables are not available, it might be considered to conduct a *reference survey*. Such a reference survey is based on a small probability sample, where data collection takes place with a mode different from the web, e.g. CAPI (Computer Assisted Personal Interviewing, with laptops) or CATI (Computer Assisted Telephone Interviewing). Under the assumption of no nonresponse, or ignorable nonresponse, this reference survey will produce unbiased estimates of the population distribution of auxiliary variables.

Using an estimated population distribution in post-stratification results in the same expected value for the estimator. So, the conditions under which the bias is reduced are the same as those for the normal 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 organisations use 'webographics' or 'psychographic' variables to divide the population in 'mentality groups'. People in the same groups are assumed to have more or less the same level of motivation and interest to participate in such surveys. If this were the case, such variables can be effectively used in weighting adjustment. This requires of course, that adequate information on psychographics is available for the population, based on high response rate random samples.

The reference survey approach also has a disadvantage. Bethlehem (2007) shows that if a reference survey is used, the variance of the post-stratification estimator is for a large part determined by the size of the small reference survey. So, the large number of observations in the online survey does not help to produce accurate estimates. One could say that the reference survey approach reduces the bias of estimates at the cost of a higher variance.

7. Sample size and response rate: the wrong paradigm?

The theoretical framework of probability sampling show that large sample sizes and high response rates will have an impact on the quality of estimates. The former will reduce sampling error which implies increased precision and smaller confidence intervals. The latter are expected to reduce nonresponse error which means an increased accuracy and thus smaller bias.

Because of the large numbers of persons in a panel or web survey it will be possible to select (in online panels) or identify (in online surveys) a sufficient number of members of a specific group (single mothers, Cape Verdeans, persons with physical handicaps, non-voters, dog-owners). Because of the large sample sizes, high response rates and presence of small subgroups, combined with shorter turnaround times and lower costs, many people argue that online self-selection surveys are better than, or at least an affordable alternative for, probability surveys that are usually smaller and that also suffer from nonresponse.

However, as mentioned above, selection probabilities in online research are generally unknown. The sampling paradigm in which sample size and response rates affect precision and accuracy of estimates is rooted in probability sampling and cannot be transferred to non-probability sampling, if only because there are many different types and models for non-probability sampling (Groves, 1989, p. 249). One could argue that the importance of sample size and response rates do not belong to the paradigm of online surveys.

Sample size

A frequent misunderstanding about online research is that large numbers make a sample better. Couper (2000) comments on the claims of a self-selected online survey: ‘“We received more than 50,000 responses – twice the minimum required for scientific validity –”..... while the survey did not yield a random sample and the selection probabilities are unknown, “this does not mean that the survey cannot yield representative social science data” (Emphasis in the original). They claim that the selection probabilities can be ‘estimated’ by comparing the distributions on standard demographic variables to official government statistics and applying weighting. This assertion is based on the assumption that matching two ‘samples’ on a variety of demographic characteristics will ensure that they also match on the survey variables of interest.’ (pp. 480-481). Not surprisingly, despite the large number of respondents, they did not resemble the U.S. population on a number of key indicators.

Dillman and Bowker (2001) express a similar opinion about online surveys: ‘Conductors of such surveys have in effect been seduced by the hope that large numbers, a traditional indicator of a high quality survey (because of low sampling error), will compensate in some undefined way for whatever coverage and nonresponse problems that might exist. Large numbers of volunteer respondents, by themselves, have no meaning. Ignoring the need to define survey populations, select probability samples, and obtain high response rates, together provide a major threat to the validity of web surveys.’

Couper (2001, pp. 173, 184) also pointed to the misguided assumption that large samples necessarily mean more valid responses. Only in the case of probability samples does an increase of sample size to an increase of precision. In non-probability samples, no inference to the underlying population is possible, and larger samples do not necessarily give better estimates than smaller samples.

Large online surveys have the advantage that specific subgroups can be identified. Information about such groups may be difficult to obtain in traditional surveys, because few people belong to these groups, they are hard to identify, or unlikely to participate in surveys. Again, the underlying assumption is that the elderly single women, low educated, ethnic minorities or other usually underrepresented groups who participate in an online survey are similar to people with the same characteristics but who do not participate. In some cases this might be a likely assumption, in others definitely not, and in most cases it will be difficult to test.

An additional caveat is that self-selection in panels and online surveys may require heavy weighting because of vastly varying participation propensities. Because of large weights the effective sample size is likely to be much smaller than the number of participants in a survey (Duffy et al., 2005). The effectiveness of a sample should be corrected by the average of the squared weights would result in a substantial reduction when even some weights are very large.

Response rate

One of the main selling points of online panels used to be high response rates. Even though response rates in online panels are rapidly decreasing (see below) this claim ignores why high response rates are important, namely because high response rates ensure that everybody who was randomly selected in the sample actually also participated. When initial recruitment is based on self-selection or when initial nonresponse is high, high response rates on a survey

may hide a wide range of other survey errors (see also Bethlehem, 2007). Sudman and Kalton (1986) discussed the use of mail panels as a means of sampling rare populations twenty years ago. Their conclusions are still valid: 'Although 80% to 90% of panel households cooperate on a study, the major problem with mail panels is that the initial cooperation rate of households invited to participate in a panel is often 10% or less. Mail panels are usually balanced by major demographic variables to remove the most obvious selection biases, but other biases still remain. These unknown selection biases may distort the survey results, and the researcher will not be able to assess the possible distortion unless some independent check can be made'.

It is possible in a panel to generate a very high response rate by approaching only those who always participate when they are invited. This is unlikely to improve survey estimates, as this over-eager group might provide highly biased results. This shows again that high response rates in a panel will generally not be able to compensate for low initial cooperation, or an unknown selection bias. Defining response rate as the response propensity of willing respondents or boosting response rates by pre-selecting the most cooperative panel members makes response rates difficult to compare with those of probability samples. For this reason, response rates in non-probability samples do not have the same meaning as those in probability samples.

Nowadays, the response rates in panels are rapidly declining. This was one of the major concerns at the 2006 ESOMAR conference on panel research in Barcelona. A comparative Belgian study, for instance, reported the following response rates: 54% (random walk, face-to-face), 21% (online panel), 15% (random sample, mail) and 12% (random sample, telephone), see Schillewaert et al. (2006). Response rates of 20% do not seem to be unusual at all now. Because of this, two additional advantages of online panels may be losing their value. Firstly, it was assumed that weighting for under-represented groups (due to sampling errors, under-coverage and non-response) would not be necessary in a sample from a panel, because the structure and composition of the panel could be determined in advance (as in quota samples). Doing this is now much more difficult because of decreasing response rates. Control over the final composition can now only be achieved by taking into account unequal response propensities from many different groups, based on information about earlier participation in similar surveys. This again assumes that non-responding panel members in a specific group are similar to respondents.

Another purported advantage of online panels was that the topic of the survey had no impact on participation, because the decision to participate in a panel is general and the response rates on individual surveys was high. This would be an advantage because one important cause of nonresponse bias is the relationship between the topic of a survey and the decision to participate. With declining response rates in online panels, interest in the topic may have become an important determinant of survey participation again. Of course, in online surveys based on self-selection, the topic of the survey is likely to be the most important determinant of participation, possibly resulting in highly biased results.

8. Time and money

The turnaround time of online research is generally much shorter than in research based on other modes of data collection. In face-to-face surveys and telephone surveys several attempts have to be made to get in touch with a sample person, and to convert initial refusers. In mail

surveys, questionnaires have to be printed and sent to target respondents, followed by reminders. Fieldwork may take weeks or even months. In online surveys – where the emphasis is more on mass than on response rates – the preparation takes less time, and fieldwork takes far less time. According to Day et al. (2006, p. 268) “... the first 12-24 hours are the most important in any online project with approximately two-thirds of panellists responding in this period”.

This short turnaround time is a very great advantage of online research, contrasting sharply with long periods and many efforts spent on contacting hard-to-contacts sample persons in face-to-face surveys (see for instance Lynn et al., 2002; Stoop, 2005) in order to enhance response rates and reduce nonresponse bias. The difference in turnaround times reflects the different aims of traditional surveys and online surveys. In the first case, ideally, a lot of efforts are done to obtain the participation of every sample member – at the cost of a long survey period and lots of money – in order to come up with correct key estimates. In the second case it is possible to collect information on a current issue within a very short time, where the focus is less on minimizing survey errors and more on the number of respondents, speed and costs. These are two different survey paradigms that are hard to compare because the aims and strategies differ. And finally, one can of course wonder who these people are who complete questionnaires within the first hours after they survey has been launched.

Online surveys and surveys among a sample from an online panel are usually much less expensive than surveys based on other modes of data collection. In the former case, most of the costs are fixed costs, supplemented with some additional costs for remuneration of respondents. There are no costs for interviewers, no costs for printing and mailing, and no costs for data processing. Indeed, the increasing costs of traditional surveys were one of the main reasons for the spectacular growth of online research. As has been shown above, however, the theoretical underpinning and the aims of traditional survey research and online research do seem to differ substantially. Costs are very important, but costs should always be seen against the background of the purpose of a survey. Deming said about this almost half a century ago (1960, p. 31): ‘cost has no measure without a measure of quality, and there is no way to appraise objectively the quality of a (quota) sample as there is with a probability sample’.

9. So when to use online panels and online surveys

We have tried to show that traditional surveys research and online survey research came from different survey traditions and have different aims. What is clear is that in official statistics, where key estimates have to be provided that are not likely to be disputed and that are not based on refutable modelling, probability sampling is to be preferred. Again going back in the history of survey research, take account of how Sudman (1966) felt about this ‘To be more explicit, where survey results will receive very sophisticated analysis or when critical decisions will be based on them, it will be worthwhile to pay a substantial cost to achieve high standards of sampling, processing, and control. Thus, the Census Bureau rightly has very high standards on the Current Population Surveys. On the other hand, many exploratory studies do not require such high standards since the analysis may be more limited and the questionnaire may itself be a major source of error. Here quota sampling would be justified.’ Deville (1991) seems to agree ‘Official statisticians, on the other hand, are responsible for data that can be used by the entire society; and that can be used, in particular, in the

arbitration of disputes between various groups, parties, and social classes. Official statistics should not tolerate any uncontrollable bias in its products. It should carry out sample surveys using probabilistic methods.'

Probability samples may have a long turnaround time, be very expensive and require substantial funding and great efforts to achieve high response rates. On the other hand, they have a secure foundation in statistical theory which allows inferences to the target population. Acceptable response rates (Stoop, 2005) are still possible. So what to do?

Recently, a number of studies have been published comparing online panels and traditional surveys (Duffy et al., 2005; Schillewaert et al., 2006). One problem is that sometimes there are differences and sometimes there are not, and generalizing across online panels or across topics is very difficult. Web surveys have attracted the interest of a wide range of academic researchers and statisticians (see *Journal of Official Statistics*, Vol.22, No.2, 2006, special issue on web surveys). Although they are attracted, they are generally very reticent when it comes to presenting online panels as a serious alternative for random samples. One promising strain of research is the recent Dutch panel comparison study mentioned above (Vonk, Van Ossenbruggen and Willems, 2006). As usual it is clear that more research, and an open mind, is needed.

When representativeness is not an issue, online panels might be a useful tool for exploratory studies, experiments, tests, and other purposes as long as undisputed results – both point estimates and the size of relationships – are not the aim of the study. When representativeness is an issue, one sensible thing to do when preparing a survey is to take account of what Groves has to say about this at the end of a study on the relationship between nonresponse rates and nonresponse bias: '*Despite low response rates, probability sampling retains the value of unbiased sampling procedures from well-defined sampling frames. Coverage error of well-defined sampling frames can be evaluated relative to a desired target population, prior to the survey being launched. Probability sampling of the frame permits use of auxiliary variables on the frame to improve the estimation from the respondent-based data. Volunteer panels lose these advantages. Low response rate probability sample surveys need to marshal the power of auxiliary variables for post-survey adjustment*' (Groves, 2006, p. 669).

A second sensible thing to do would be to consider building a panel based on random sampling and including the Internet and the non-internet population, see Scherpenzeel (2006). An access panel conforming to strict methodological specifications could function as a host survey for a wide range of cross-sectional and longitudinal studies.

And finally, the final sensible thing to do when using non-probability online panels is to follow Fowler's admonition (2002, p. 56) and be transparent: 'If a researcher decides to use a non-probability sample, however, readers should be told how the sample was drawn, the fact that it likely is biased in the direction of availability and willingness to be interviewed, and that the normal assumptions for calculating sampling errors do not apply. Such warnings to readers are not common. In many cases, non-probability samples are misrepresented seriously, and that constitutes a serious problem for the credibility of social science research.'

References

- Bethlehem, J.G. (1988), Reduction of the nonresponse bias through regression estimation. *Journal of Official Statistics* 4, pp. 251-260.

- Bethlehem, J.G. (2002), Weighting Nonresponse Adjustments Based on Auxiliary Information. In: Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (eds): *Survey Nonresponse*. Wiley, New York.
- Bethlehem, J.G. (2007) *Reducing the bias of web survey based estimates*. Discussion Paper 07001, Statistics Netherlands, Voorburg, The Netherlands
- Bowley A.L. (1906): Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society* 69, pp. 548-557.
- Converse, J.M. (1987) *Survey Research in the United States. Roots & Emergence 1890 – 1960*. Berkeley, University of California Press.
- Couper, M.P. (2000) Web Surveys. A Review of Issues and Approaches. *Public Opinion Quarterly*, Vol. 64, pp. 464-494.
- Couper, M.P. (2001) The Promises and Perils of Web Surveys. In: Andrew Westlake, Wendy Sykes, Tony Manners and Malcom Riggs (eds.) *The Challenge of the Internet*. Proceedings of the ASC International Conference on Survey Research Methods. Chesham, UK, May 2001.
- Day, D., R. Risk, J. Koo and B. Martin (2006) *Ensuring Data Integrity for Business Decisions. An In-depth Analysis of the Components that Affect Data Quality*. Panel Research 2006. ESOMAR Publication Series, Volume 317, pp. 253-269.
- Deming, W. (1960) *Sample design in business research*. New York: Wiley.
- Deville, J.C. (1991) A Theory of Quota Surveys. *Survey Methodology*. Vol. 17, N° 2. pp. 163-181.
- Dillman, D A. Bowker, D. (2001), The web questionnaire challenge to survey methodologists. In: Reips, U.D. and Bosnjak, M. (eds.), *Dimensions of Internet Science*, Pabst Science Publishers, Lengerich, Germany.
- Duffy, B., Terhanian, G., Bremer, J. and Smith, K. (2005) Comparing data from online and face-to-face surveys. *International Journal of Market Research*, Vol. 47, No. 6, pp.615-639
- Fowler, Floyd J. Jr. (2002) *Survey Research Methods*, 3rd Edition. Thousand Oaks, California: Sage Publications.
- Groves, R.M. (1989) *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- Groves, R.M. (2006) Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, Vol. 70, pp. 646-675.
- Kalton, Graham (1987) *Introduction to Survey Sampling*. California: Sage Publications.
- Horvitz, D.G. en D.J. Thompson (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, pp. 663-685.
- Kalton, G. (1983), *Introduction to Survey sampling*. Saga Publications, Inc, Beverly Hills, California.
- Kish, L. (2003) The Hundred Years' War of Survey Sampling (Reprinted from *Statistics in Transition*, 1995). In: Graham Kalton and Steven Heeringa (eds.) *Leslie Kish, Selected Papers*. New York: Wiley, pp. 5-19.

- Kruskal, William and Frederick Mosteller (1979a) Representative Sampling, I: Non-scientific Literature. *International Statistical Review* 47, pp. 13-24.
- Kruskal, William and Frederick Mosteller (1979b) Representative Sampling, II: Scientific Literature, Excluding Statistics. *International Statistical Review* 47, pp. 111-127.
- Kruskal, William and Frederick Mosteller (1979b) Representative Sampling, III: the Current Statistical Literature. *International Statistical Review* 47, blz. 245-265.
- Lynn, Peter, Paul Clarke, Jean Martin and Patrick Sturgis (2002) The Effects of Extended Interviewer Efforts on Nonresponse Bias. In: Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J.A. Little (eds.) *Survey Nonresponse*. New York: Wiley, pp. 135-148.
- Morton-Williams, J. (1993) *Interviewer Approaches*. Aldershot: Dartmouth Publishing.
- Neyman, J. (1934): On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97, pp. 558-606.
- Saris, W.E. (1998), Ten Years of Interviewing Without Interviewers: The Telepanel. In: Couper, M.P., Baker, R.P., Bethlehem, J.G., Clark, C.Z.F., Martin, J., Nicholls II, W.L. and O'Reilly, J.M. (eds.), *Computer Assisted Survey Information Collection*. Wiley, New York.
- Scherpenzeel, Annette (2006) *Een online panel als platform voor multidisciplinair onderzoek*. Paper presented at the DANS-symposium Access panels en online onderzoek, panacee of slangenkuil.
- Schillewaert, Niels, Annelies Verhaeghe, Bert Weijters and Kristof de Wolf (2006) *Social class and life style differences between modes of data collection*. Panel Research 2006. ESOMAR Publication Series, Volume 317, pp. 174-193.
- Schnell, Rainer (1997) *Nonresponse in Bevölkerungsumfragen, Ausmaß, Entwicklung und Ursachen*. Opladen: Leske und Budrich.
- Stoop, I.A.L. (2005) *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office.
- Sudman, Seymour (1966) Probability Sampling with Quotas. *Journal of the American Statistical Association*, Vol. 61, pp. 749-771.
- Sudman, Seymour, and Graham Kalton (1986) New Developments in the Sampling of Special Populations. *Annual Review of Sociology*, Vol. 12, pp. 401-429.
- Thomas, Randall K. (2006), Online Panels: the Research Revolution. *Panel Research 2006, ESOMAR World Research*, ESOMAR Publication Services, Vol. 317, pp. 6-7.
- Vonk, T., Van Ossenbruggen, R. and Willems, P. (2006), The effects of panel recruitment and management on research results, a study among 19 online panels. *Panel Research 2006, ESOMAR World Research*, ESOMAR Publication Services, Vol. 317, pp. 79-99.

About the Authors

Jelke Bethlehem is Senior Methodologist at the Division of Methodology and Quality of Statistics Netherlands. He carries out research in the fields of survey methodology, nonresponse, computer-assisted interviewing and online surveys. He is also parttime professor in Statistical Information Processing at the University of Amsterdam. He can be contacted at Statistics Netherlands, PO Box 4000, 2270 JM Voorburg, The Netherlands, tel. +31 70 337 4995, e-mail jbtm@cbs.nl.

Ineke Stoop is head of the department of ICT and Data Services at the Netherlands' Social and Cultural Planning Office. She is responsible for the surveys sponsored by the SCP and research on data quality. She has specialised in nonresponse issues. She is also a member of the Central Coordinating Team of the European Social Survey. She can be contacted at SCP, PO Box 16164, 2500 BD, The Hague, The Netherlands, tel. +31 70 3407671, fax +31 70 3407044, e-mail i.stoop@scp.nl.

