New developments in survey data collection methodology for official statistics

10

Jelke Bethlehem

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2008–2009	= 2008 to 2009 inclusive
2008/2009	= average of 2008 up to and including 2009
2008/'09	= crop year, financial year, school year etc. beginning in 2008 and ending in 2009
2006/'07–2008/'09	= crop year, financial year, etc. 2006/'07 to 2008/'09 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

New developments in survey data collection methodology for official statistics

Jelke Bethlehem

Summary: There is a growing demand for statistical information in society. National statistical institutes have to satisfy this demand. The way they attempt to accomplish this, changes over time. Changes in survey methodology for official statistics may have been caused by new developments, for example in computer technology. Changes may also be due to new challenges like increasing nonresponse rates, decreasing budgets, or demands for reducing the response burden. This paper describes some new developments in survey methodology that may help to solve problems of survey taking in official statistics. The R-indicator is described as an additional indicator for survey quality. Web surveys are considered as a cheaper means of data collection, either as a single-mode survey or as one of the modes in a mixed-mode survey. Also attention is paid to more flexible ways of conducting the fieldwork of a survey. The R-indicator could play a role in this.

Keywords: Representativity, R-indicator, Web survey, Mixed-mode survey, Responsive design

There is an ever growing demand for statistical information in society. National statistical institutes have to satisfy this demand. The way they attempt to accomplish this, changes over time. Changes in survey methodology for official statistics may have been caused by new developments, for example in computer technology. Changes may also be due to new challenges faced by national statistical institutes, like increasing nonresponse rates, decreasing budgets, or demands for reducing the response burden.

Chapter 2 of this paper describes some historical developments. It shows how important the probability sampling paradigm has been for the current state of survey methodology as applied in official statistics. Also, the rapid developments in computer technology have had an important impact the way survey data is being collected, processed, analysed and published.

The current situation is characterized on the one hand by problems like decreasing quality (e.g. due to nonresponse and measurement errors) and limited budgets, and on the other hand, the new opportunities offered by the Internet.

Nonresponse affects the representativity of survey data, and therefore the quality of survey outcomes. Chapter 3 describes a new indicator (the R-indicator) that measures the representativity of survey response. Such an indicator can be a useful additional indicator for survey quality. It may be applied during the fieldwork of the survey to monitor data collection efforts. It may also be useful to compare a survey over time, or to compare different surveys.

National statistical offices have to produce reliable and accurate statistics. They often conduct face-to-face or telephone surveys to collect the data that form the basis for these statistics. This is an expensive way of survey data collection, but experience has shown that it is necessary in order to obtain high quality data. Now that many these offices are faces with reduced budgets, they are looking for less costly means of data collection. An alternative may be offered by web surveys. This type of survey becomes increasingly popular, but also has its methodological drawbacks. Some methodological aspects are described in chapter 4. This leads to the question whether web surveys can be used in official statistics, whether as a single mode survey, or as one of the modes in a mixed-mode survey. Chapter 5 attempts to find an answer to this question. This chapter also considers a mixed-mode survey as a means to reduce nonresponse rates. Response behaviour may depend on the data collection mode. By approaching people with the mode most fit for them, they may be more inclined to respond.

Chapter 6 is about responsive survey design. Its objective is to make the fieldwork more effective by splitting it into a number of phases. The implementation of the next phase depends on the results of the previous phase. This could mean focusing fieldwork on a specific group with a specific mode. Such an approach requires information on the progress of the fieldwork. These so-called paradata play an important role in modelling and implementing such an approach.

1. Some history

1.1 The emergence of probability sampling

The idea of conducting surveys for compiling statistical overviews of the state of affairs in a country is already very old. 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. All these surveys were complete enumerations of the population. The idea of sampling had not yet emerged.

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 1000 and 1500 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 idea of using sampling instead of a complete enumeration came up around the year 1895. In that year, Anders Kiaer (1895, 1997), the founder and first director of Statistics Norway, published his Representative Method. He proposed questioning only a (large) sample of persons who together formed a 'miniature' of the population. 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.

For a number of years, there were two methods of sample selection. The first one was Kiaer's Representative Method, based on purposive (non-probability) 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 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 principles of probability sampling form the foundation of modern survey research. They are vital for making valid inference about the population being investigated. These principles have been successfully applied in official and academic statistics since the 1940's, and to a much lesser extent also in commercial market research. The message is that if samples are not based on probability sampling, it is not possible to compute unbiased estimates, and it is also not possible to quantify their margins of error. Moreover, probability sampling allows for a clearer trade-off between quality and costs.

1.2 The impact of computer technology

The practical instruments for data collection also have changed over the years. Until the 1970's, paper questionnaire forms were used, either in face-to-face surveys, telephone surveys or mail surveys. The rapid developments in computer technology changed that. It started with Computer Assisted Telephone Interviewing (CATI). The first CATI systems were developed by commercial market research agencies in the

United States. These systems not only could handle interviewing by telephone from a centralised facility, but also took care of call scheduling and case management. See Nicholls and Groves (1986) for an overview. CATI was followed by Computer Assisted Personal Interviewing (CAPI). This is face-to-face interviewing by interviewers using a laptop to ask the questions and record the answers. CAPI emerged in the 1980's when lightweight laptop computers made face-to-face interviewing with a computer feasible. European national statistical institutes played an important role in these developments. Early CAPI experiments are described in CBS (1987). Self-administered forms of CAI also emerged during the 1980's. It is called in Computer Assisted Self Interviewing (CASI) or Computer Assisted Self-Administered Questionnaires (CASAQ). The electronic questionnaire runs on a computer in the respondent's home. After having completed the questionnaire, the respondents sends the data to the statistical agency. More on CASI and related techniques can be found in Couper et al. (1998).

More recently, particularly in commercial market research, face-to-face, mail and telephone surveys are increasingly replaced by web surveys. The popularity of web surveys is not surprising. Since many people have access to the Internet, a web 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 web surveys offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation and movies).

Of course, the increased use of computers is not limited to just data collection. One example is the development of powerful software tools for data editing, and another is the introduction of dynamic graphics for publishing statistics on the Internet.

2. The quest for representativity

2.1 The response rate as a quality indicator?

Most surveys suffer from nonresponse. This is the phenomenon that sample elements do not provide the required information. Nonresponse may seriously affect the quality of the survey outcomes. Estimates of population characteristics will be biased if, due to non-response, some groups in the population are over- or underrepresented, and these groups behave differently with respect to the survey variables.

Nonresponse is a serious problem in many countries. As an example, figure 2.1.1 shows the response rates in the first round of the European Social Survey. To keep country results comparable, the sampling design was similar in all countries. Target was to realise a response rate of at least 70%. Figure 2.1.1 shows that some countries did not manage to achieve this. The graph also shows there are large differences in

response rates between countries. They vary from 33% in Switzerland to 80% in Greece. A more extensive analysis of the nonresponse in the ESS can be found in Stoop (2005).

Figure 2.1.1. Response rates in the European Social Survey
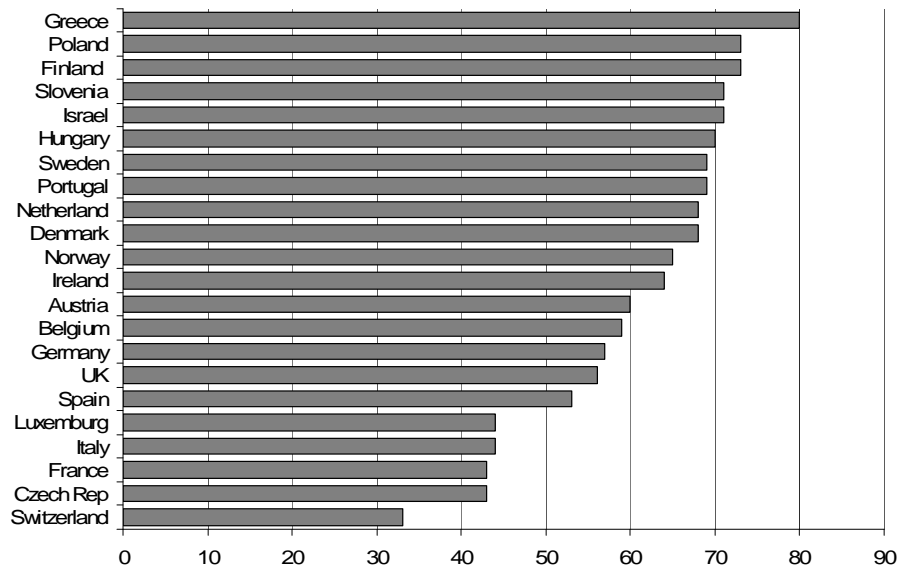
The question is now whether one should conclude that the survey results in Switzerland are much more unreliable than those in Greece? The answer is not necessarily yes. Survey agencies often use the survey response rate as an indicator of survey quality. However, a low response rate does not necessarily has to lead to inaccurate survey estimates. If non-response is completely random, i.e. there is no correlation between response behaviour and the survey variables, estimates will still be unbiased. Indeed, the literature on survey methodology contains ample examples showing that response rates by themselves are poor indicators of non-response bias. As an indicator of survey quality it can be misleading.

This is illustrated by an example using data from an anonymised public use data file containing data from a survey of Statistics Netherlands. This survey will be called the General Population Survey (GPS) in this paper. Fieldwork covered a period of two months. The mode of data collection in the first month was CAPI. Non-respondents were approached in the second month by CATI if they had a listed, land-line phone. Otherwise, CAPI was used again. Table 2.1.1 contains estimates of two population quantities: the percentage of people receiving a social allowance and the percentage of non-natives. Both variables are taken from a register and are artificially treated as survey questions. Therefore percentages for the complete sample are also available. These sample percentages are given in table 2.1.1.

After one month of fieldwork the response rate was 46.4%, while after the full two month period the rate had increased to 58.7%. So the second month of fieldwork increased the response by 12.3%. This did, however, not result in better estimates. The bias of the estimators increased after the second month in both cases.

Table 2.1.1. Response means in the GPS after
the first and second month of data collection

Variable	After 1 month	After 2 months	Sample
Social allowance	11.4 %	11.2 %	12.9 %
Non-native	12.4 %	12.1 %	14.7 %
Response rate	46.4 %	58.7 %	100.0 %

There is a need for additional survey quality indicators providing more insight in the possible risk of biased estimators. This section describes such an indicator. It is called the *R-indicator*. The *R* stands for 'representativity'. The R-indicator measures how representative the survey response is, or to say it differently, how the composition of the response differs from that of the sample.

The R-indicator can be used in many different ways. One way is to inspect the survey data after completion of the fieldwork. It can also play an important role during data collection. By monitoring the fieldwork, data collection efforts can be targeted at obtaining a response the composition of which does not deviate too much from that of the complete sample (or the population).

2.2 What is representativity?

The concept of *representativity* is often used in survey research, but usually it is not clear what it means. 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 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. They recommended not using the word *representative*, but instead to specify what one means.

To be able to define an indicator for representativity, the concept of representativity is defined here as the absence of selective forces. Every element k in the population is assumed to have a certain, unknown, probability ρ_k of responding when selected in the sample. It is clear that there are no selective forces if all response probabilities are equal. Unfortunately, response probabilities are unknown in practice. Therefore they have to be estimated using the available data. To this end, the concept of response propensity is introduced. The *response propensity* of element k is defined by

$$\rho(X_k) = P(R_k = 1 \mid X_k),$$ (2.2.1)

where $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})'$ is the vector of values of p auxiliary variables for element k. The response indicator R_k assumes the value 1 if element k is selected in the sample and responds; otherwise R_k assumes the value 0. So the response propensity is the probability of response given the values of some auxiliary variables. The response propensities are also unknown, but they can be estimated

provided the values of the auxiliary variables are available for both the respondents and non-respondents. To be able to estimate the response propensities, a model must be chosen. The most frequently used one is the logistic regression model. It assumes the relationship between response propensity and auxiliary variables can be written as

$$\text{logit}\left(\rho(X_k)\right) = \log\left(\frac{\rho(X_k)}{1-\rho(X_k)}\right) = \sum_{j=1}^{p} X_{kj}\beta_j \qquad (2.2.2)$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is a vector of regression coefficients. The logit transformation ensures that estimated response propensities are always in the interval [0, 1].

2.3 The R-indicator

The R-indicator measures how far the composition of the survey response deviates from the original sample. If all response probabilities are equal, the response is representative, and there will be no systematic differences between the composition of the response and the sample. If the response probabilities are not equal, it is important to establish to what extent the composition of the response is affected. This is accomplished by defining a distance function that measures how far the individual response probabilities differ from the mean response probability.

Suppose, that the individual response probabilities $\rho_1, \rho_2, \dots, \rho_N$ of all elements in the population are known. Then the standard deviation

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(\rho_k - \overline{\rho})^2} \; , \qquad (2.3.1)$$

of the response probabilities can be computed. This is a distance function, namely the Euclidean distance. $S(\rho) = 0$ if all response probabilities are equal, and the value of $S(\rho)$ will be larger as there is more variation in the values of the response probabilities. One can prove that the maximum value of $S(\rho)$ is equal to 0.5. The R-indicator is now defined as

$$R(\rho) = 1 - 2S(\rho) \qquad (2.3.2)$$

This R-indicator assumes a value in the interval [0, 1]. A value of 1 implies strong representativity. The smaller its value is, the more the response composition deviates from that of the sample composition, and the less representative the response is.

The values of the individual response probabilities are unknown in practice. This is solved by estimating response probabilities as defined in (2.2.1), for example with a logit model. Usually, the type of available auxiliary information does not allow for computing estimated response probabilities for all elements in the population, but just for the sample elements, If the estimated probabilities in the sample are denoted by $\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_n$, then the R-indicator (2.3.2) can be estimated by

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n}\frac{(\hat{\rho}_k - \overline{\hat{\rho}})^2}{\pi_i}} \; , \qquad (2.3.3)$$

where π_i is the first order inclusion probability of sample element i. Note that expression (2.3.3) involves two estimation steps. The first one is estimation of the response probabilities and the second one is estimation of the standard deviation.

The R-indicator has been applied in a large scale follow-up study among the non-respondents in the Dutch Labour Force Survey (LFS) in 2005. Two samples of non-respondents were approached once more using either a call-back approach with the full LFS questionnaire or a basic-question approach with a very short questionnaire containing only a few basic questions. Some results are summarised in table 2.3.1. For more details see Schouten (2007) and Cobben and Schouten (2007). The R-indicator was estimated using a logistic regression model including a large number of explanatory variables that measure demographic, geographic and socio-economic characteristics of the households.

Table 2.3.1. Comparing R-indicators in the LFS follow-up study

Response	Response rate	R-indicator
LFS	62.2 %	0.80
LFS + call-back approach	76.9 %	0.85
LFS + basic-question approach	75.6 %	0.78

The value of the R-indicator for the initial LFS response is equal to 0.80, which is lower than the ideal value of 1.00. So this response is not completely representative. Application of the call-back approach increases the response rate from 62.2% to 76.9%. The value of the R-indicator also increases, from 0.80 to 0.85. This indicates that the additional response improves the composition of the data set. Application of the basic-question approach results in a different conclusion. Although the response rate increases from 62.2% to 75.6%, the value of the R-indicator drops from 0.80 to 0.78. Apparently, the basic-question approach does not improve the composition of the data set. This approach gives 'more of the same' and, hence, sharpens the contrast between respondents and non-respondents.

Another example shows the effect of the use of incentives in an attempt to increase the response rate of the Dutch Labour Force Survey. The sample was randomly split into three groups. People in the first group did not receive an incentive. For people in the second group a stamp book with five stamps was included in the pre-notification letter. People in the third group received 10 stamps. The results of the experiment are summarized in table 2.3.2.

Table 2.3.2. The impact of incentives on representatvity

Response	Response rate	R-indicator
No incentive	66.6 %	0.86
5 stamps	72.2 %	0.82
10 stamps	73.8 %	0.84

It is clear that incentives have a positive effect on the response rate. Giving 5 stamps increases the response rate from 66.6% to 72.2%. And for 10 stamps the response rate goes up to 73.8%. Unfortunately, the composition of the response does not improve. The value of the R-indicator is even somewhat lower. Apparently, incentives do not help to improve the response of specific groups in the population. It turned out in this experiment that particularly for non-natives incentives did not work.

2.4 Use of the R-indicator

The R-indicators can be used in several different ways in the survey process. A number of possibilities are described here:

- Monitoring the survey process. It may already become clear during data collection that the composition of the collected data differs from that of the initial sample. This may lead to a decision to initiate additional efforts to obtain data for specific groups in the target population.

- Controlling the survey process. Use of an R-indicator while collecting data may reveal that the composition of the response is deviating more and more from representativity. This could lead to a decision to drastically change the survey design for the remainder of the data collection process. For example, a different data collection mode could be implemented. This mid-fieldwork decision to change the design is called "responsive survey design". See also section 5.

- Pre-selection of auxiliary variables for non-response correction. Estimation of response probabilities is based on models involving auxiliary variables. Variables that significantly contribute to predicting response probabilities are also important in non-response correction techniques like adjustment weighting.

- Analysis of surveys. The R-indicator can be used as a simple analysis tool providing insight in possible problems due to non-response. Like the response rate, it is a quality indicator. The R-indicator can also be very useful for comparing surveys over time or comparing survey data for different domains or regions.

If the R-indicator is used in making decisions for additional data collection efforts for specific groups, one should realize that this may affect the selection probabilities of elements in the population. These effects have to be accounted for in constructing unbiased estimators of population parameters.

Another aspect of the R-indicator is that its value increases as the values of the response propensities decrease. This may seem counter-intuitive, but it reflects the notion that a small, but representative sample is more reliable than a large, but unrepresentative, sample. In addition, it is not advocated to use the R-indicator as one single indicator for survey, but to use it in combination with other indicators like the response rate.

The R-indicator proposed in this paper is promising because it can be estimated using sample data and it allows for easy interpretation. Computation of its value is

reasonably straightforward with standard statistical software packages like SPSS, SAS or STATA. If the R-indicator is to be used for monitoring or controlling the survey process, the data collection system used must be able to compute the R-indicator 'on the fly'.

Research with respect to the R-indicator is still in progress. One of the issues is the comparability of R-indicators for different surveys or of the comparability of R-indicators for the same survey over time. The value of the R-indicator is partly determined by the auxiliary variables used in the logit model for the response propensities. If the model contains too few auxiliary variables, there will be less variability in the estimated response propensities and therefore the R-indicator may be biased upwards. This calls for a bias correction term.

It is the objective of the RISQ project to develop and to test representativity indicators. RISQ stands for *Representativity Indicators for Survey Quality*. Five partners participate in this project: Statistics Netherlands, Statistics Norway, The Statistical Office of Slovenia, the University of Southampton (UK) and the University of Leuven (Belgium). The RISQ project is financed by the 7th Framework Programme of the European Union. More information can be found on www.risq-project.eu.

Research with respect to the R-indicators will focus on the statistical properties of estimators of R-indicators, the circumstances under which these indicators can be computed, and how these indicators should be applied in survey research. A particular challenge is to include the use such indicators in fieldwork operations, as they may help to increase both the cost-effectiveness of survey data collection and the quality of the survey response.

3. The conquest of the web

3.1 Web surveys

Web surveys have become increasingly popular over the last couple of years. This is not surprising. A web 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. Web surveys also offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation and movies).

At first sight, web surveys have much in common with other types of surveys. It is just another mode of data collection. Questions are not asked face-to-ace or by telephone, but over the Internet. There are, however, major methodological issues. One issue is under-coverage. Since data are collected using the Internet, people without Internet access will never be able to participate in a web survey. This means

research results only apply to the Internet population and not to the complete population.

Another issue is that sample selection is often based on self-selection of respondents instead of on probability sampling. The principles of probability sampling have not been applied. Researchers have no control over the selection mechanism, resulting in unknown selection probabilities. Therefore, no unbiased estimates can be computed, nor can the accuracy of estimates be established.

A third issue is measurement errors. Many surveys in official statistics are CAPI or CATI surveys. These are not the cheapest modes of data collection, but they are used because response rates are high and data quality tends to be good. What would change in this respect if a CAPI or CATI survey was to be replaced by a web survey, where no interviewer is present?

These three problems (under-coverage, self-selection and measurement) are discussed below in some more detail.

3.2 Under-coverage

Web surveys suffer from under-coverage because the target population is usually much wider than just the Internet population. According to data from Eurostat, 54% of the households in the EU had access to Internet in 2007. There were large variations between countries. The countries with the highest percentages of Internet access were The Netherlands (83%), Sweden (79%) and Denmark (78%). Internet access was lowest in Bulgaria (19%), Romania (22%) and Greece (25%). For more information, see Eurostat (2007).

Even more problematic is that Internet access is unevenly distributed over the population. A typical pattern found in many countries is that elderly, low-educated and ethnic minorities are severely under-represented among those having access to Internet. Bethlehem (2009) shows that the bias of the response mean as an estimator of the population mean of a variable Y is equal to

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}),$$ (4.2.1)

where the subscript I denotes the Internet population and NI the non-Internet population. 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. Therefore the bias decreases as Internet coverage increases. The second factor is the contrast $\bar{Y}_I - \bar{Y}_{NI}$ between the means of 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. It is not clear whether the contrast between those with and without Internet 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.

If under-coverage in a web survey really is a problem, a possible solution could be to simply provide Internet access to those without Internet. An example of this approach is the LISS panel, 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. Co-operative households without Internet access were provided with equipment giving them access to Internet.

It should be noted that the problem of under-coverage is not unique for web surveys. For example, telephone surveys suffer more and more from under-coverage, because less and less people have a listed telephone number.

3.3 Self-selection

If respondents of a web survey are recruited by means of self-selection, estimates will be biased. Self-selection means that the survey is simply put on the web. Participation requires in the first place that respondents are aware of the existence of the survey. They have to accidentally visit the website, or they have to follow up a banner, e-mail message, or a call in another commercial. 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 population has unknown probability ρ_k of participating in the survey. Bethlehem (2009) shows that the expected value of the sample mean is equal to

$$E(\bar{y}) \approx \bar{Y}^* = \frac{1}{N} \sum_{k=1}^{N} \frac{r_k}{\bar{r}} Y_k \qquad (3.3.1)$$

where $\bar{r}$ is the mean of all response propensities. The bias of this estimator is equal to

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{R_{rY} S_r S_Y}{\bar{r}}, \qquad (3.3.2)$$

in which $R_{\rho Y}$ is the correlation coefficient of the target variable and the response probabilities, S_ρ is the standard deviation of the response probabilities, and S_Y is the standard deviation of the target variable. It can be shown that in the worst case (S_ρ assumes its maximum value and the correlation $R_{\rho Y}$ is equal to either +1 or -1) the absolute value of the bias is equal to

$$\left| B_{\max}(\bar{y}) \right| = S_Y \sqrt{\frac{1}{\bar{r}} - 1}. \qquad (3.3.3)$$

Bethlehem (1988) shows the formula (3.3.2) also applies in the situation in which a probability sample has been drawn, and subsequently nonresponse occurs during the fieldwork. Consequently, expression (3.3.3) provides a means to compare potential biases in various survey designs. For example, regular surveys of Statistics Netherlands are all based on probability sampling. Their response rates are around 70%. This means the absolute maximum bias is equal to 0.65 ´ S_y. One of the largest self-selection web surveys in The Netherlands was *21minuten.nl*. Within a period of

six weeks in 2006 about 170,000 people completed the web questionnaire. The target population of this survey was not defined, as everyone could participate. If it is assumed the target population consists of all Dutch from the age of 18, the average response propensity is equal to 170,000 / 12,800,000 = 0.0133. Hence, the absolute maximum bias is equal to $8.61 \times S_y$. It can be concluded that the bias of the large web survey can be a factor 13 larger than the bias of the smaller probability survey. Apparently, by taking a random sample, the bandwidth of a possible bias is reduced.

The effects of self-selection can also be illustrated using an example related to the general elections in The Netherlands in 2006. Various survey organizations used opinion polls to predict the outcome of these elections. The results of these polls are summarized in table 3.3.1. Differences of two seats or more are printed in boldface. *Politieke Barometer*, *Peil.nl* and *De Stemming* are opinion polls carried out by market research agencies. They are all based on samples from web panels. To reduce a possible bias, adjustment weighting was carried out. The polls were conducted one day before the election. The Mean Absolute Difference indicates how big the differences (on average) are between the poll and the election results. Deviations of 3 seats and more are underlined. Particularly, differences are large for the more volatile parties like PvdA, SP and the PVV. For example, one poll predicted 32 seats in parliament for the SP (socialist party) whereas this party in fact got only 25 seats.

DPES is the Dutch Parliamentary Election Study. The fieldwork was carried out by Statistics Netherlands in a few weeks just before the elections. The principles of probability sampling were followed here. A true (two-stage) probability sample was drawn from the population register. Respondents were interviewed face-to-face (using CAPI). The predictions of this survey were much better than those based on the online opinion polls. The predictions and election results only differ for four parties, and differences are at most one seat.

Table 3.3.1. Parliamentary elections in The Netherlands (2006), predictions and results

	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	**<u>37</u>**	**<u>38</u>**	31	32
VVD (Liberals)	22	23	22	21	22
SP (Socialist)	25	23	23	**<u>32</u>**	26
GL (Green party)	7	7	8	5	7
D66 (Liberal democrats)	3	3	2	1	3
ChristenUnie (Christian)	6	6	6	8	6
SGP (Christian)	2	2	2	1	2
PvdD (Animal party)	2	2	1	2	2
PVV (Populist)	9	**<u>4</u>**	**<u>5</u>**	**<u>6</u>**	8
Other parties	0	2	1	2	1
Mean absolute difference		1.27	1.45	2.00	0.36

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 a poll on a website. People could vote for one of the nominated books or mention another book of their own choice. More than 90,000 people participated in the survey. The winner turned out to be the new 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 be considered to be representative of the Dutch population.

Can self-selection web surveys be used for data collection in official statistics? The discussion in this section leads to the conclusion that severe methodological problems make it very hard, if not impossible, to make valid inference about the population that is surveyed. Self-selection can cause estimates of population characteristics to be biased. This seems to be similar to the effect of nonresponse in traditional probability sampling based surveys. However, it was shown that the bias in self-selection surveys can be substantially larger.

Self-selection is a serious problem, but it can be solved by applying probability sampling. A random sample (e.g. of addresses) can be drawn from a sampling frame. A letter can be sent to each selected address with request to complete a questionnaire on the Internet. Unique identification codes guarantee that the proper persons answer the questions. In fact, the only difference with a mail questionnaire is that the paper questionnaire form is replaced by an electronic one on the Internet.

The LISS panel is an interesting example of a web panel where the problems of under-coverage and self-selection have been addressed. This panel was constructed by means of a true probability sample. 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 of probability surveys than to those of self-selection web surveys.

3.4 Measurement errors

With respect to data collection, there is a substantial difference between CAPI and CATI on the one hand and web surveys on the other. Interviewers carry out the fieldwork in a CAPI or CATI survey. There are no interviewers, however, in a web survey. It is a self-administered survey. Therefore quality of collected data may be lower due to higher nonresponse rates and more errors in answering questions.

De Leeuw (2008) and Dillman et al. (2008) discuss differences between various modes of data collection. They observe that a positive effect of the presence of

interviewers is that they can assist respondents in getting the right answers to the questions. Interviewers can motivate respondents, answers questions for clarification, provide additional information and remove causes for misunderstanding.

The presence of interviewers can also have a negative effect. It will lead to more socially desirable answers for questions about potentially sensitive topics. There is also a tendency to agree more with statements made in questions if interviewers are present (acquiescence). Without interviewers, respondents may feel more anonymous, and therefore will be more inclined to answer sensitive questions honestly.

Another aspect is the way in which the questions are offered. In case of CAPI or CATI, questions are read out loud by the interviewers. Respondents listen to the questions and answer them. The interviewers can check that questions are well-understood, and they can assist respondents in determining the proper answer. In case of a web survey, the respondents have to read the questions themselves. There is no guarantee that they read the questions carefully and that they understand them. If they do not understand a question, they tend to transform it into another question that they do understand. So, an answer is given to a different question.

The mode of data collection may affect the way a closed question is answered, particularly if the list of answer options is long. There seems to be a tendency to select the last option in the list when the options are read out loud (recency effect), whereas respondents seem to have a preference for the first option in the list when they have to read the list themselves (primacy effect).

CAPI and CATI are both a form of computer assisted interviewing (CAI). The interviewers enter the answers of the respondents after which the interviewing system determines the next question to be asked. The software system is completely in control of navigation through the questionnaire. It is not possible for respondents to jump to questions of their own choice. Most web surveys do not have built-in navigation. Usually respondents are free to jump back and forth through the questionnaire. This makes it possible to skip certain questions or change one's opinion about the answer to a question.

It may happen that respondents do not know the answer to a question. They lack the information to correctly answer a factual question or they really may not have an opinion on a specific issue. This suggests it must always be possible to offer "don't know" as one of the answer options. There are several ways to do this:

1) Do not offer "don't know" as a possible answer. This forces respondents to give a 'real' answer whether or not it is the true answer. This may cause errors in answers and irritation among respondents. As a consequence, respondents may even break off the interview. This is probably more likely in web surveys, as there are no interviewers to persuade respondents to continue. Couper (2008) notes that a forced answer also violates the norms of voluntary survey participation.

2) Offer "don't know" as an answer option, but not explicitly. Particularly in CAPI/CATI software, the option "Don't know" is often not mentioned. If, however, respondents indicate they do not know the answer, the interviewer can record this by means of a special key combination. This possibility cannot be implemented in exactly the same way in a web survey.

3) Offer "don't know" explicitly as an answer option. Research by e.g. Kalton et al. (1978) and Tiemeijer (2008) indicates that people tend to avoid this option because it is socially undesirable to have no opinion. The problem is also described by Bishop (2004) and Couper (2008). This effect is probably stronger for interviewer assisted surveys. For other types of questions the "don't know" can be an easy way out for the respondents, because they do not have to think about the answer.

4) Reduce the embarrassing effect of not knowing the answer by introducing a filter question. This question asks whether the respondents have an opinion. Only if they say they have, they are asked what it is. This generally leads to a higher percentage of "don't know", and this may be closer to the truth.

CAPI and CATI have the advantage that some form of error checking can be implemented. It means that answers to questions are checked for consistency. Errors can be detected during the interview, and therefore also corrected during the interview. It has been shown (see e.g. Couper et al., 1998) that this can improve the quality of the collected data. The question is now whether error checking should be implemented in a web survey? What happens when respondents are confronted with error messages? Maybe they just correct their mistakes, but it may also happen that they will become annoyed and stop answering questions.

Error messages in CAPI/CATI can be complex, involving the answers to several questions. Resolving the situation may involve going back to earlier questions and changing their answers. This may be a too complex operation for many respondents. It seems to be possible, however, to implement simpler checks and error messages. Examples are checks on whether an answer was entered in a field, or whether a date has a proper format. Couper (2008) advises that error messages should at least be polite, illuminating and helpful, and certainly should not blame the respondents for the detected problem.

In the end, dealing with errors and error messages may be a trade-off between non-response and data quality. Further research should make clear what the best approach is.

The length of the questionnaire is a final issue to be mentioned here. If a questionnaire is too long respondents may refuse to participate, or they may stop somewhere in the middle of the questionnaire. Questionnaires of CAPI surveys can be longer than those of CATI en web surveys. It is more difficult to stop a face-to-face conversation with an interviewer than to hang up the phone or to stop somewhere in the middle of web survey questionnaire. Literature seems to suggest that CATI interviews should not last longer than 50 minutes, and completing a web survey questionnaire should not take more than 15 minutes.

3.5 Mixed-mode 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 national statistical institutes 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-mode approaches. The first approach is the use of different modes 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 rates 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. Mode effects occur if the same question produces a different answer when asked in a different mode. The presence or absence of interviewers may be a source of mode effects. The presence of interviewers leads to more socially desirable answers, particularly for questions about potentially embarrassing behaviour. The presence of interviewers also causes acquiescence. This is the tendency to agree with statements by interviewers. It is easier to agree than to disagree.

The interviewers are in control of the presentation of questions to respondents in CAPI and CATI surveys. They can see to it that the respondents hear and understand every word of it. When necessary, additional explanation can be provided. This is different for self-completion surveys. There is no guarantee that questions are carefully read and clearly understood.

There are also mode effects with respect to answering closed questions. Section 4.4 already described the primacy effect for mail and web surveys, and the recency effect for CAPI and CATI surveys. Also, the treatment of "don't know" may lead to mode effects.

There are two approaches to reduce mode effects. One is to develop separate questionnaires for different modes. A specific question may be defined differently in different modes as long as it measures the same thing. The different versions of the question should be cognitively equivalent. This is not very easy to realise, as it may take substantial research and experimentation.

Dillman (2008) proposes his so-called *unimode* approach. This is a set of guidelines to define questions in such a way that the mode effects are minimized. Here are some examples:

- Keep all answer options the same across modes.
- Include all answer options in the text of the question.
- Reduce the number of answer options as much as possible.
- Reverse the order of the answer options in half of the questionnaires.
- Develop equivalent instructions for skip patterns

Instead of reversing the order of the answer options in half of the questionnaires one could also think of randomizing the order of the answer options. It may be difficult, or even impossible, to implement equivalent skip instructions for paper questionnaires.

One may wonder whether it is possible to develop a questionnaire which completely satisfies all unimode guidelines. Particularly of attitudinal questions, it may turn out to be necessary to define mode-dependent versions.

A final aspect of mixed-mode surveys to be mentioned here is case management. This is of vital importance, particularly in case of a sequential mixed-mode approach. A case management system should see to it that cases are assigned to the proper mode at the proper moment. Cases may not disappear from the system. Also, duplicate cases must be avoided. This calls for a sophisticated overall case management system.

3.6 Can web surveys be used in official statistics?

Can web surveys be used in official statistics, were focus is on obtaining precise and unbiased estimates of population characteristics? The previous sections described a number of potential problems. This section explores whether these problems can be solved.

It was already shown in section 4.2 that web surveys suffer from under-coverage. Since there are differences between those with and without Internet access, under-coverage will often cause estimates to be biased. Fortunately, Internet penetration will increase over time. This helps to reduce the bias. However, it is possible 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.

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 e.g. 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.

Many web surveys rely on self-selection of respondents. It was shown in this paper that self-selection can cause estimators to be substantially biased. This makes self-selection surveys useless for official statistics. Application of the principles of probability sampling is of crucial importance. Without it, no unbiased estimators can be computed, nor can margins of errors be determined.

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 all 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 by phone and asking them for their e-mail address. If they provide an e-mail address, they are sent a link to the online questionnaire form.

It should be noted that also surveys based on probability sampling have their problems. One of the most important ones is probably nonresponse. In fact, this also introduces a form of self-selection. Fortunately, it can be shown that the bandwidth of a potential bias is much smaller than in web surveys.

Another way to reduce the under-coverage problem is to conduct some form of mixed-mode survey. Groups in the population with low Internet penetration can be approached with a different mode of data collection. As a simplified example, young people could be approached with a web mode and the elderly with a CAPI mode.

In contemplating a change from a CAPI or CATI survey to a web survey, measurement errors must be taken into account. It was shown there are all kinds of mode effects: If the same question is asked in a different mode, the answer will be different. A change to a web survey will not always decrease data quality. For example, sensitive questions will be answered better if the there are no interviewers. On the other hand, quality may also decrease because of the lack of assistance of interviewers.

Some problems with mode effects may be solved by formatting questions in a different way, using Dillman's guidelines for unimode questionnaires. Other issues need further research. One is the use of error messages in web surveys. Another is the treatment of "don't know". Vis-Visschers et al. (2008) describe a small experiment at Statistics Netherlands that seems to suggest that "don't know" should always be included as one of the answers options for factual questions. For

attitudinal questions it may be better to include filter questions asking whether the respondents have an opinion.

Response rates are an important issue. Response rates for traditional CAPI and CATI surveys vary between 60% and 70% at Statistics Netherlands. The first experiences with web surveys (based on probability sampling) result in response rates of around 30%. See, for example Beukenhorst & Wetzels (2009). And an experiment with a housing demand survey showed that for large and complex questionnaires response rates may even drop to 20%.

4. Fixed or flexible?

4.1 Responsive survey design

Statistical agencies in many countries experience decreasing response rates in surveys. There seems to be a growing reluctance to participate in surveys. Efforts to keep response at an acceptable level increase survey costs. Decreasing response rates also affect the quality of survey results. There is a higher risk of biased estimates of population characteristics. This trend calls for a new approach. Groves & Heeringa (2006) propose an approach called *responsive survey design*.

Traditional survey designs are fixed. Aspects like sampling design, mode of data collection, respondent recruitment protocols, number of call-backs, and use of incentives are all decided upon in the design phase, and never changed. However, it may turn out during the data collection process that these decisions are not the best ones, and that changes are required in order to obtain more reliable and more accurate statistics. This is the idea behind responsive survey design.

As an illustration, table 4.1.1 repeats the results of the Dutch GPS survey. It contains estimates of two population quantities: the percentage of people receiving a social allowance and the percentage of non-natives. Since both variables are taken from a register, their sample percentages are known. CAPI was used in the first month of the fieldwork. Non-respondents with a listed phone number were approached in the second month by CATI. CAPI was used again for other nonrespondents. The table shows the estimates deteriorated in the second month of fieldwork. The question is whether better results could have been obtained if fieldwork decisions for the second month are based on results in the first month.

Table 4.1.1. Response means in the GPS
after the first and second month of data collection

Variable	After 1 month	After 2 months	Sample
Social allowance	11.4 %	11.2 %	12.9 %
Non-native	12.4 %	12.1 %	14.7 %
Response rate	46.4 %	58.7 %	100.0 %

According to Groves & Heeringa (2006), the fieldwork of the survey is assumed to consist of a number of so-called design phases. All survey design features (e.g. sampling design, mode of data collection, recruitment procedures) remain fixed during a design phase. The survey design features can be different in each phase. For example, the mode of data collection in the first phase could be CAPI, whereas a different one could be used in subsequent phases. Responsive survey design consists of the following four steps:

- Identify a set of survey design features that may affect costs and quality of the survey. Typical examples of such features are the mode of data collection, sample size, and number of call-back attempts;

- Define a set of indicators to measure (during data collection) the survey design features identified in step 1. Typical indicators are the response rate, the R-indicator described in section 3.3, and interviewer costs.

- Measure the cost and quality indicators in each phase. Decide at the end of a phase to change the design features of the next phase based on the values of the indicators.

- At the end of the fieldwork, combine the data collected in the separate design phases to obtain single estimators for population characteristics.

So the difference with traditional survey design is that during the fieldwork decisions may be taken to change the fieldwork based on information obtained during the fieldwork. This approach resembles to some extent the ideas about quality control that were proposed by Deming (1986) in his famous book on improving quality and productivity in industry. Many of his famous 14 points for management also apply to the production of statistical information. One of these points states that one should cease dependence on mass inspection. Inspection of the final product to improve quality is too late, ineffective and costly. Quality must be built in at the design stage. This also applies to data collection. By trying to detect and correct problems after the fieldwork has been completed, one fails to locate problems immediately after they have occurred. Consequently, these problems are not solved.

If a specific survey is conducted repeatedly, information from previous surveys can be used to optimise the current design. Such a strategy could be called differential survey design or adaptive survey design. Such an approach is particularly interesting for survey agencies if the can link survey data to register data.

4.2 Using response probabilities

One way to implement responsive survey design is to use response probabilities. Suppose this means that each element k in the population has unknown probability ρ_k of participating in a survey design phase. Bethlehem (2009) shows that the expected value of the sample mean is equal to

$$E(\bar{y}) \approx \bar{Y}^* = \frac{1}{N} \sum_{k=1}^{N} \frac{r_k}{\bar{r}} Y_k \qquad \text{(4.2.1)}$$

where $\bar{r}$ is the mean of all response propensities. See also expression (3.3.2). The bias of this estimator is equal to

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{R_{rY} S_r S_Y}{\bar{r}}, \qquad \text{(4.2.1)}$$

in which $R_{\rho Y}$ is the correlation coefficient between the target variable and the response probabilities, S_ρ is the standard deviation of the response probabilities, and S_Y is the standard deviation of the target variable. The bias is small if the variation of the response probabilities is small.

The R-indicator measures the variation of the response probabilities. A value close to 1 means the probabilities are approximately equal. There will be no large bias. A small value of the R-indicator implies the variation of the response probabilities is substantial. There will be groups with low response probabilities. The next survey design phase should focus on these groups. Special treatment may increase the response probabilities.

As an example, table 4.2.1 contains the standard deviation of the estimated response probabilities and the R-indicator after the first and after the second month of fieldwork of the GPS. Apparently, the response after the first month is not completely representative, as the R-indicator has a value of 0.803.

Table 4.2.1. The R-indicator in the GPS after the first and second month of data collection

| Variable	Standard deviation of response probabilities	R-indicator
After 1 month	0.099	0.803
After 2 months	0.117	0.766

The ideas of response survey design were not applied in this survey. The survey design features for the second phase (the second month) were fixed in advance. It is clear that the situation became worse in the second month. The value of the R-indicator reduced to 0.766.

Table 4.2.2. Auxiliary variables in logit model

Variable	Description	Categories
Gender	Gender	2
Married	Is married	2
Age13	Age in 13 age groups	13
Ethnic	Type of non-native	5
HHSize	Size of the household	5
HHType	Type of household	5
Phone	Has listed phone number	2
Hasjob	Has a job	2
Region	Region of the country	5
Urban	Degree of urbanization	5

To explore what went wrong in the first phase, response probabilities were estimated. A logit model was used. To find such a model, auxiliary variables are required. These variables must have been measured for both respondents and nonrespondents. Since the GPS survey data file could be linked to the Social Statistics Database (SSD) of Statistics Netherlands, many such variables were available. Table 4.2.2 contains the set of variables that turned out to have a significant contribution to the model. The distribution of the estimated response probabilities is displayed in figure 4.2.1. The histogram shows substantial variation. The lowest probability is 0.112 and the highest one is 0.665.

Figure 4.2.1. The distribution of estimated response probabilities in the first month of the GPS.
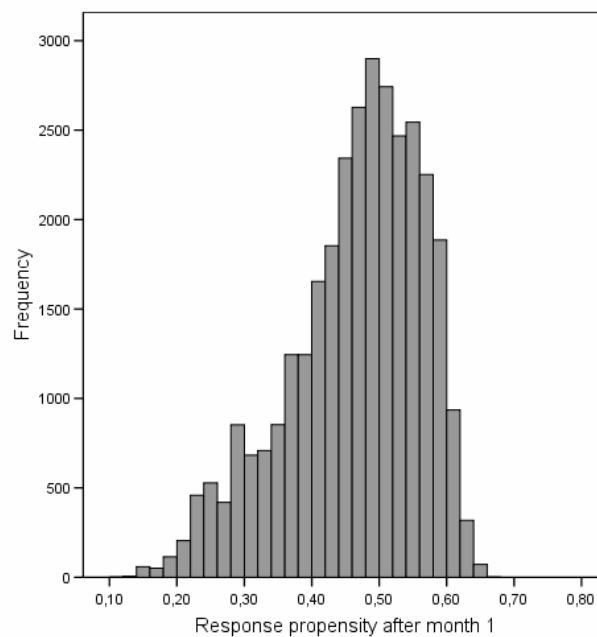
To be able to implement an effective survey design for the second phase, insight must be obtained into which groups have a low response probability. This requires analysis of the response probabilities in relation to the auxiliary variables. As an illustration, table 4.2.3 shows the characteristics of the persons with the lowest and highest response probabilities.

The table shows that unemployed non-natives in the big cities are under-represented in the response after the first month. Furthermore, natives in rural areas are over-represented. To reduce the variation in response probabilities after the second phase, it could be a good idea to focus on getting more response from the unemployed non-natives in urban areas. Little effort should go into getting response from natives in rural areas.

Table 4.2.3. The lowest and highest response probability

Variable	Value for lowest propensity	Value for highest propensity
Gender	Male	Female
Is married	No	Yes
Age in 13 age groups	45 - 49	70 – 74
Type of non-native	First generation non-western	Native
Size of the household	2	4
Type of household	Other	Couple without children
Has listed phone number	No	Yes
Has a job	No	No
Region of the country	Metropolis	Greenfields
Degree of urbanization	Very strong	Not
Response propensity	0.112	0.665

4.3 Response survey design in practice

A responsive survey design can only be implemented if a sufficient amount of information about the data collection process becomes available during the fieldwork of the survey. In the era of paper questionnaire forms this was difficult to accomplish, but computer-assisted interviewing has made it possible to record a wealth of information about the data collection process. This type of information is usually called *paradata*. See Couper (1988). These paradata should at least contain information about each contact attempt in the survey (date, time, result, etc).

Computer assisted interviewing systems like Blaise are capable of producing paradata. Laflamme (2009) described an application at Statistics Canada were the focus is on reducing survey costs.

Tools have to be developed implementing the cost and quality indicators discussed above. Examples of such indicators are the response rate and the R-indicator. There may also be a need for cost indicators.

Of course, the survey data collection organization as a whole has to be capable of implementing responsive survey design. It means that a fixed approach in which everything is planned beforehand, has to be replaced by a more flexible approach.

5. References

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

Bethlehem, J.G. (2009), *Applied Survey Methods – A Statistical perspective.* John Wiley & Sons, Hoboken, NJ,, USA.

Beukenhorst, D. & Wetzels, W. (2009), A comparison of two mixed-mode deisgns of the Dutch Safety Monitor: mode effects, costs and logistics. Report DMH-2009-07-14-BKT, Statistics Netherlands, Methodology Dpartment, Heerlen, The Netherlands.

Bishop, G.F. (2004). *The Illusion of Public Opinion: Fact and Artifact in American Public Opinion Polls*. Rowman and Littlefield Publishers, Lanham, MD, USA

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.

Bowley, A.L. (1926): Measurement of the precision attained in sampling. Bulletin of the International Statistical Institute, XII, Book 1, pp. 6-62.

CBS (1987), Automation in Survey Processing, CBS Select. Staatsuitgeverij, The Hague, The Netherlands..

Cobben, F. (2004), *Nonresponse correction techniques in household surveys at Statistics Netherlands: a CAPI-CATI comparison*. Technical report, Statistical Netherlands, Methods and Informatics Department, Voorburg, The Netherlands.

Cobben, F. & Schouten, B. (2007), A follow-up with basic questions of nonrespondents to the Dutch Labour Force Survey. Discussion paper 07011, Statistics Netherlands, Voorburg, The Netherlands.

Couper, M.P. (1998), Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 41-49.

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

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*. Wiley, New York, USA.

De Leeuw, E.D.(2005), To mix or not to mix data collection modes in surveys. *Journal of Official Statistics,* Vol. 21, No. 2, pp. 233 – 255.

De Leeuw, E.D. (2008), Choosing the mode of data collection. In: De Leeuw, E.D., Hox, J.J. & Dillam, D.A. (eds.), *International Handbook of Survey Methodology*, Lawrence Erlbaum Associates, New York.

Deming, W.E. (1986), *Out of the Crisis*. Massachusetts Institute of Technology, Cambridge, MA. USA.

Dillman, D.A., Smyth, J.D. & Christian, L.M. (2008), *Internet, Mail, and Mixed-Mode Surveys, The Tailored Design Method*. John Wiley & Sons, Hoboken, USA.

Eurostat (2007), *More than 40% of households have broadband internet access*. Eurostat News Release 166 / 2007. Eurostat, Luxembourg.

Groves, R.M. and Heeringa, S.G. (2006), Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society: Series A, 169, pp. 439 – 457.

Kalton, G., Collins, M. & Brook, L. (1978), Experiments in Wording Opinion Questions. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 27, pp. 149-161.

Kiaer, A. N. (1895), Observations et expériences concernant des dénombrements représentatives. Bulletin of the International Statistical Institute, IX, Book 2, pp. 176-183.

Kiaer, A. N. (1997 reprint): Den repräsentative undersökelsesmetode. Christiania Videnskabsselskabets Skrifter. II. Historiskfilosofiske klasse, Nr 4 (1897). English translation: The Representative Method of Statistical Surveys, Statistics Norway.

Kruskal, W. & Mosteller, F. (1979a), Representative sampling I: non-scientific literature. International Statistical Review 47, pp. 13-24.

Kruskal, W. & Mosteller, F. (1979b), Representative sampling II: scientific literature excluding statistics. International Statistical Review 47, pp. 111-123.

Kruskal, W. & Mosteller, F. (1979c), Representative sampling III: current statistical literature, International Statistical Review 47, pp. 245-265.

Kuusela, V. (2003), Mobile phones and telephone survey methods. In R. Banks, J. Currall, J. Francis, L. Gerrard, R. Kahn, T. Macer, M. Rigg, E. Ross, S. Taylor & A. Westlake (eds.), *ASC 2003 - The impact of new technology on the survey process. Proceedings of the 4th ASC international conference,* pp. 317-327. Association for Survey Computing (ASC), Chesham Bucks, UK.

Laflamme, F. (2009), *Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada.* Invited Paper, 57th Session of the International Statistical Institute, Durban, South Africa.

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.

Nicholls, W.L. & Groves, R.M. (1986), The status of computer assisted telephone interviewing. Journal of Official Statistics 2, pp. 93-134.

Scherpenzeel, A. (2008), An online panel as a platform for multi-disciplinary research. In: I. Stoop & M. Wittenberg (eds.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam, pp. 101-106.

Scherpenzeel, A. & Bethlehem, J. (2010), How representative are online-panels? Problems of coverage and selection and possible solutions. In: M. Das, P. Ester, L. Kaczmirek & P. Mohler (eds.), *Social Research and the Internet: Advances in applied Methods and New Research Strategies*, Routlegde Academic, New York, USA. To be published.

Schouten, B. (2007), A follow-up of nonresponse in the Dutch Labour Force Survey. Discussion paper 07004, Statistics Netherlands, Voorburg, The Netherlands.

Stoop, I.A.L. (2005), The Hunt for the Last Respondent: Nonresponse in Sample Surveys. Social and Cultural Planning Agency, The Hague, The Netherlands.

Tiemeijer, W.L. (2008), *Wat 93,7 Procent van de Nederlanders moet weten over Opiniepeilingen*. Aksant, Amsterdam, The Netherlands.

Vis-Visschers, R., Arends-Tóth, J., Giesen, D. & Meertens, V. (2008), *Het aanbieden van 'weet niet' en toelichtingen in een webvragenlijst*. Report DMH-2008-02-21-RVCS, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands.