# The Future of Surveys for Official Statistics

*Jelke Bethlehem, Statistics Netherlands, Methodology Department*

## 1. The ever changing landscape of survey research

### 1.1 The role of Blaise in the world 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 (PAPI) to computer-assisted interviewing (CAI). It started in the 1970's with Computer Assisted Telephone Interviewing (CATI). At that time the first CATI systems were developed by commercial market research agencies in the United States. CATI systems could not only handle interviewing by telephone from a centralised facility, but also took care of call scheduling and case management. CATI was followed by Computer Assisted Personal Interviewing (CAPI). CAPI emerged in the 1980's when lightweight laptop computers made face-to-face interviewing with a computer feasible. Self-administered forms of CAI also emerged during the 1980's. It was sometimes called in Computer Assisted Self Interviewing (CASI). The electronic questionnaire ran on a computer in the respondent's home. Respondents completed the questionnaire on their computer at home, after which the data were transmitted to the statistical agency. And now, particularly in commercial market research, face-to-face, mail and telephone surveys are increasingly replaced by web surveys. The popularity of Computer Assisted Web Interviewing (CAWI) is not surprising. Now that many people have Internet, a web survey is a simple means to get access to a large group of people.

The development of the Blaise System has followed the trends in survey research. Development of the system started in 1980. The first version came out in 1986. Blaise 1.0 was a CADI system. The basic idea was to create systems that could improve the handling of paper questionnaire forms by integrating data entry and data editing tasks. It ran under the operating system MS-DOS.

After version 1 of Blaise had been in use for a while, it was realised that the system could be made much more powerful by implementing computer assisted interviewing (CAI). Version 2.0 of Blaise was completed in 1988. It implemented just one form of computer assisted interviewing (CAPI). In 1990, another mode of computer assisted interviewing was included: Computer Assisted Telephone Interviewing (CATI). This was version 2.3 of Blaise.

When the operating system MS-DOS was gradually replaced by Windows, this had consequences for Blaise. Version 4 of Blaise marked the change from MS-DOS to Windows. Blaise 4 was the first Windows version of Blaise. It was released in 1998. The functionality of Blaise 4 was basically the same as that of the previous MS-DOS version. Of course, the graphical user interface offered much more possibilities for screen layout.

In the first 10 years of its existence, the basic concept of Blaise, the definition of the questionnaire in the Blaise language, turned out to be sufficiently powerful to cope with all developments in information technology and survey methodology.

Then came the rise of the Internet. As more and more people were connected to the Internet, computer assisted web interviewing (CAWI) gained in popularity amongst researchers as means of data collection. CAWI was implemented in Blaise 4.6. It was released in 2003. And again, the basic concept could cope with this new data collection technique without dramatic changes.

What will be the future developments in survey research? And what is needed in Blaise to keep up with these developments. This paper attempts to answer this question by pointing at some of these developments. After describing some history and future challenges of survey methodology, the rest

of the paper concentrates on two issues. The first one is mixed-mode data collection and the role web surveys can play in this. The second one is the collect data of acceptable quality under the constraint of a reduced survey budget. This may requires a new approach (such as responsive survey design) and new tools (such as the R-indicator).

### 1.2 Some history of survey research

The idea of conducting surveys for compiling statistical overviews of the state of affairs in a country is already very old, see Kendall (1960). 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. The Roman Empire 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.

*Figure 1.2.1. RAPI: Rope Assisted Personal Interviewing*

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, see figure 1.2.1. 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 were questioned 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.

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

The principles of probability sampling formed to basis for modern survey taking. 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 more commercial market research. Where samples are not based on probability sampling, it is not possible to compute unbiased estimates, and it is also not possible to quantify margins of error.

What has changed over the years, are the instruments for actual data collection. Until the 1970's, all surveys used paper forms (PAPI) in either face-to-face, telephone or mail surveys. Then it became possible used computers for data collection. The paper form was replaced by a desktop or laptop computer. It simplified the work of the interviewers, produced higher quality data, and substantially reduced to time needed to carry out a survey. After the 1990's more and more people were connected to the Internet, and web surveys became possible. Conducting a survey was even more simpler, faster and cheaper.

## 1.2 Future trends

National statistical institutes in many countries are faced with three conflicting developments: (1) they face substantial budget constraints, (2) they are confronted with a demands for more and more detailed information, and (3) the have to reduce the response burden of surveys as much as possible. And of course, the quality of the published statistical information should remain at an acceptable level. Not surprisingly, many statistical are rethinking their data collection operations.

This paper explores some possible future directions for survey research to go. The first one is an increased use of web surveys. This is not without risks. Opportunities and threats are discussed in the first part of section 2. Another possible direction is to used web surveys as one of the modes in mixed-mode surveys. Several statistical organizations are already using mixed-mode survey or are planning to introduce mixed-mode surveys. Some of the challenges and also the implications for Blaise are discussed in in the second part of section 2.

Reducing survey costs and response burden would mean collecting as little data as possible, while maintaining an acceptable level of survey quality. This calls for adequate indicators of survey quality. It is argued in section 3 that the response rate is insufficient for this. A different indicator is proposed: the R-indicator. This indicator can also be used during data collection to monitor the fieldwork and make changes in the survey when necessary. This is called responsive survey design.

# 2. The conquest of the web

## 2.1 Single-mode 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. And web surveys offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation and movies).

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. 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 can only apply to the Internet population and not to the complete population. Another issues is that sample selection is often based on self-selection of respondents instead of on probability sampling. 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. These problems are discussed below in some more detail.

Web surveys suffer from under-coverage because the target population is usually much wider than just the Internet population. According to data from Eurostat, the statistical office of the European Union, 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%).

Even more problematic is that Internet access is unevenly distributed over the population. A typical pattern found in many countries is that the elderly, the low-educated and ethnic minorities are severely under-represented among those having access to Internet. Bethlehem (2007) shows that the bias of the response mean as an estimator of the population mean of a variable $Y$ is equal to

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}), \qquad (2.1.1)$$

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, however, whether the contrast between the 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.

Application of self-selection will also cause estimates to 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 a 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 $\rho_k$ of participating in the survey. Bethlehem (2008) shows that the expected value of the sample mean is equal to

$$E(\bar{y}) \approx \bar{Y}^* = \frac{1}{N}\sum_{k=1}^{N}\frac{\rho_k}{\bar{\rho}}Y_k \qquad (2.1.2)$$

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

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{R_{\rho Y}S_\rho S_Y}{\bar{\rho}}, \qquad (2.1.3)$$

in which $R_{\rho Y}$ is the correlation coefficient of the target variable and the response probabilities, $S_\rho$ 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 ($S_\rho$ assumes it maximum value and the correlation $R_{\rho Y}$ is equal to either +1 or -1) the absolute value of the bias is equal to

$$\left|B_{\max}(\bar{y})\right| = S_Y\sqrt{\frac{1}{\bar{\rho}} - 1}\,. \qquad (2.1.4)$$

Bethlehem (1988) shows the formula (2.1.3) also applies in the situation in which a probability sample has been drawn, and subsequently nonresponse occurs during the fieldwork. Consequently, expression (2.1.4) 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 vary are around 70%. This means the absolute maximum bias is equal to $0.65 \times S_{y.}$ 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 he bias of the smaller probability survey.

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 2.1.1. Differences of two seats or more are printed in boldface. *Politieke Barometer*, *Peil.nl* and *De Stemming* were 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. Particularly, differences are large for the more volatile parties like PvdA, SP and the PVV. Differences of three seats or more are printed in boldface. For example, one poll predicted 32 seats in parliament for the SP (socialist party) whereas this party got only 25 seats. This difference is larger than could be explained by the sampling error.

Table 2.1.1. 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 | 37 | 38 | 31 | 32 |

| | | | | | |
|---|---|---|---|---|---|
| VVD (liberals) | 22 | 23 | 22 | 21 | 22 |
| SP (socialist) | 25 | 23 | 23 | 32 | 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 | 4 | 5 | 6 | 8 |
| Other parties | 0 | 2 | 1 | 2 | 1 |
| Mean absolute difference | | 1.27 | 1.45 | 2.00 | 0.36 |

*DPES* is the Dutch Parliamentary Election Study. The fieldwork was carried out by Statistics Netherlands in a few weeks just before the elections. The probability sampling principle has been 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 are 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.

Probability sampling has the additional advantage that it provides protection against certain groups in the population attempting to manipulate the outcomes of the survey. This may typically play a role in opinion polls. Self-selection does not have this safeguard. An example of this effect could be observed in the election of the 2005 Book of the Year Award (Dutch: NS Publieksprijs), a high-profile literary prize. The winning book was determined by means of a poll on a website. People could vote for one of the nominated books or mention another book of their 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 t very hard, if not impossible, to draw valid conclusions about the population to be 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 letters 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 problem of under-coverage in web surveys has to be addressed too. It is interesting to note that only between 60% and 70% of the households in The Netherlands still have a listed landline telephone. So one out of three households is missing if a sample is selected from a telephone directory. This shows that also more traditional phone surveys suffer from under-coverage. It is expected that under-coverage for web surveys will decrease over time. From the point of view of coverage, a web survey may be better than a telephone survey, at least in The Netherlands.

If under-coverage in web survey really is a problem, a possible solution could be to simple 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. So sample selection was based on true probability sampling. Moreover, co-operative households without Internet access were provided with equipment giving them access to Internet. Analysis by Scherpenzeel & Bethlehem (2010) shows that the results of this panel are closer to those of surveys based on probability sampling than to those based on self-selection web surveys.

Can a web survey be an alternative for a CAPI or CATI survey? With respect to data collection, there are substantial differences between CAPI and CATI on the one hand and web surveys on the other. Interviewers carry out the fieldwork for CAPI and CATI surveys. They are important in convincing people to participate in the survey, and they also can assist in completing the questionnaire. There are no interviewers 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 the answers to the questions. However, response to sensitive questions may be higher and better without interviewers.

## 2.2 Mixed-mode surveys

Budget cuts on the one hand and demands for more and more detailed information on the other, while maintaining an acceptable level of data quality, have stimulated statistical agencies 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 to use different modes concurrently. The sample is divided into groups and each group is approached by a different mode. The other approach is to use different modes sequentially. All sample persons are approached by one mode. The non-respondents are then followed up by a different mode than the one used in the first approach. This process can be repeated for a number of modes.

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

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

Mixed-mode survey 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 presenting the questions to the 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-completing surveys. There is on guarantee that questions are carefully read and clearly understood.

There are also mode effects with respect to answering closed questions. Research seems to suggest that respondents in mail or web surveys more often choose the first answer option (primacy effect), while there is a preference for the last answer option in CAPI and CATI surveys (recency effect).

A final mode effect to be mentioned here is caused by the treatment of “don’t know”. This option is often not offered explicitly in CAPI or CATI surveys, but the interviewers have a facility to record this answer if the respondents insist they really do not know the answer. For self-administered surveys, this option is either explicitly offered or it is not possible at all to give this answer. If ‘don’t know” is clearly one of the answer options, more respondents will select this option ( the easy way out).

It will be clear that a mixed-mode survey may suffer from mode-effects. There are two approaches to reduce these 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. Moreover, it may turn out to be cumbersome and error prone to maintain three different Blaise questionnaires for one survey.

Dillman (2008) proposed 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 of guidelines:

- 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 questionnaire 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. This may lead to one Blaise questionnaire, but with rule instructions like

```
Question_A
IF Mode = CAPI THEN
   Question_B1
ELSEIF Mode = CATI THEN
   Question_B2
ELSEIF Mode = CAWI THEN
   Question_B3
ENDIF
Question_C
```

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 an overall case management system.

# 3. The quest for representativity

## 3.1 The response rate as a quality indicator?

Most surveys suffer from non-response. This is the phenomenon that sample elements do not provide the required information. Non-response may seriously affect the quality of the outcomes of a survey. 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.

Survey agencies often use the survey response rate as an indicator of survey quality. However, a low response rate does not necessarily imply that the accuracy of survey estimates is poor. If non-response is ignorable, i.e. there is no direct 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 from the 1998 Dutch POLS survey (short for Permanent Onderzoek Leefsituatie or Integrated Survey on Household Living Conditions in English). Table 3.1.1 contains estimates of two population quantities: the percentage of people receiving some form of social allowance and the percentage of people having at least one parent that was born outside the Netherlands. The people are, by definition, non-native. Both variables are taken from a register and are artificially treated as survey questions. Therefore sample percentages are also available. These sample percentages are given in table 3.1.1. After one month of fieldwork the response rate was 47.2%, while after the full two month period the rate had increased to 59.7%. The mode of data collection in the first month was CAPI (Computer Assisted Personal Interviewing). Non-respondents were approached in the second month with CATI (Computer Assisted Telephone Interviewing) if they had a listed, land-line phone. Otherwise, CAPI was used again. The second month of fieldwork increased the response by 12.5% This did, however, not result in better estimates. The bias of the estimators increased after the second month.

Table 3.1.1. Response means in POLS after the first and second month of data collection

| Variable | After 1 month | After 2 months | Sample |
|---|---|---|---|
| Social allowance | 10.5 % | 10.4 % | 12.1 % |
| Non-native | 12.9 % | 12.5 % | 15.0 % |
| Response rate | 47.2 % | 59.7 % | 100.0 % |

There is a need for additional survey quality indicators that provide 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’. R-indicators measure how representative the survey response is, or to say it differently, how the composition of the response differs from that of the sample.

R-indicators can be used in many different ways. One way is to inspect the survey data after completion of the fieldwork. But they 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).

## 3.2 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 sta-

tistics 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 $\rho_k$ 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 the response propensity is introduced. The *response propensity* of element $k$ is defined by

$$\rho_k(X) = P(R_k = 1 | X_k), \tag{3.1.1}$$

where $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})'$ is a vector of values of auxiliary variables. 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

$$logit(\rho_k(X)) = log\left(\frac{\rho_k(X)}{1-\rho_k(X)}\right) = \sum_{j=1}^{p} X_{kj}\beta_j, \tag{3.1.2}$$

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

### 3.3 The R-indicator

The R-indicator measures how far the composition of the response to a survey 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 $\rho_1, \rho_2, \ldots, \rho_N$ of all elements in the population are known. Then the standard deviation

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(\rho_k - \bar{\rho})^2}, \tag{3.3.1}$$

of the response probabilities can be computed. This is a distance function. $S(\rho) = 0$ 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 be defined as

$$R(\rho) = 1 - 2S(\rho) \tag{3.3.2}$$

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 the sample composition.

The values of the individual response probabilities are unknown in practice. This is solved by estimating response propensities as defined in (3.1.1), for example with a logit model. Estimates $\hat{\rho}_1, \hat{\rho}_2, ..., \hat{\rho}_n$ can only be computed for the sample elements. The R-indicator (3.3.2) can now be estimated by

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n}\frac{(\hat{\rho}_k - \hat{\bar{\rho}})^2}{\pi_i}}, \qquad (3.3.3)$$

where $\pi_i$ is the first order inclusion probability of sample element *i*. Note that expression (3.3.3) 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 was 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 3.3.1. For more details see Schouten (2007) and Cobben and Schouten (2007). The R-indicator was estimated using logistic regression models including a large number of explanatory variables that measured demographic, geographic and socio-economic characteristics of the households.

Table 3.3.1. 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.

## 3.4 Use of the R-indicator

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

- Monitoring the survey process. Already during data collection it can become clear whether or not the composition of the collected data differs from that of the initial sample. The outcomes of this monitoring process may help to underpin 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 already during the data collection phase may reveal that the composition of the collected data may deviate more and more from representativity. This could lead to a decision to focus the remainder of the data collection process on groups that are under-represented. Groves & Heeringa (2006) call these mid-survey decisions to change the design "responsive survey design". See also section 3.5.

- 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 times or comparing survey data for different domains, regions or countries.

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 software 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. It is the objective of the RISQ project to develop and to test the R-indicator. RISQ stands for Representativity Indicators for Survey Quality. Five partners participate in this European 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.r-indicator.eu.

### 3.5 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 increases 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 are all decided 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 reliable statistics. This is the idea behind response survey design.

According to Groves & Heeringa (2006), the fieldwork of the survey is assumed to consist of a number of phases. 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 could be used in subsequent phases. Responsive survey design consists of the following four steps:

1) Identify survey design features that potentially 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;

2) 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 3.4, and interviewer costs.

3) 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.

4) 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. These statements particularly apply to data collection. By trying to detect and correct problems in the fieldwork well after they have occurred, one fails to locate the source of these problems, and consequently, these problems can not be solved.

A responsive survey design can only be implemented of a sufficient amount of information about the data collection process becomes available during the data collection process. Fortunately, computer assisted interviewing systems like Blaise are capable to produce these so-called paradata. Tools have to be developed that implement the cost and quality indicators discussed above.

Of course, the survey data collection as a whole has to be capable of implementing responsive survey design. Again this requires a case management system that is sufficiently powerful to transfer cases between data collection modes.

More about the topic of responsive design can be found in Groves & Heeringa (2006), Mohl & Laflamme (2007) and Wagner & Raghunathan (2007).

## 4. Some conclusions

To prepare Blaise for the future in official statistics, attention should be paid developments in survey methodology. These developments are partly triggered by increasing nonresponse rates and demands for reduction of costs and response burden. These developments are visible world-wide.

Even in the early days of Blaise, the system was already presented as an integrated system for survey processing. For example, Bethlehem (1990) stressed the importance of consistency between data collections modes that was enforced by Blaise. This makes it very suitable for implementing mixed-mode surveys with the Dillman's unimode approach. What is lacking, is a set of tools or system to implement an effective case management systems. Of course, it is possible to develop a tailor-made system for each separate mixed-mode survey, but that requires substantial efforts and resources. A generalized case management system should recognize that different survey organizations may have different requirements. Therefore it may be a challenge to develop a standard system. Blaise already has a CATI call management system. Possible they experiences with this system could help.

The literature on survey methodology research shows there is a demand for tools to monitor the data collection process. The R-indicator is such a tool. And more are needed to implement responsive survey design. Such tools should be able to access Blaise data and paradata files. It is to

be expected that some of these tools require complex computations, like maximum likelihood estimation. It may turn out that a tool like Manipula is not implement these indicators. Of course, it is always possible to develop tools as independent programs, like Bascula. To give survey researchers the possibility to experiment with new tools, it would be nice to have a direct link between Blaise and, for example, the R language. R is an open source language and environment for statistical computing and graphics. Many powerful statistical techniques have already been implement in R, see www.r-project.org.

## 5. References


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

Bethlehem, J.G. (1990), The Blaise System for Integrated Survey Processing. CBS-report, Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

Bethlehem, J.G. (2007), *Reducing the bias of web survey based estimates*. Discussion Paper 07001, Statistics Netherlands, Voorburg/Heerlen, The Netherlands.

Bethlehem, J.G. (2008), *How accurate are self-selection web surveys?* Discussion Paper 08014, Statistics Netherlands, The Hague/Heerlen, 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.

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. & 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., 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*. John Wiley & Sons, 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.

Deming, W.E. (1986), Out of the crisis. Cambridge University Press, Camebridge.

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.

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.

Kendall, M.G. (1960), Where shall the history of statistics begin? *Biometrika*, 47, pp. 447-449.

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.

Mohl, C. & Laflamme, F. (2007), Research and responsive design options for survey data collection at Statistics Canada. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp.2962-2968.

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.

Wagner, J. & Raghunathan, T. (2007), Bayesian approaches to sequential selection of survey design protocols. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp.3333-3340.