Chapter One

The Road to Web Surveys

1.1 Introduction

Web surveys are a next step in the evolution process of survey data collection. Collecting data for compiling statistical overviews is already very old, almost as old as mankind. All through history, rulers of countries used statistics to take informed decisions. However, new developments in society always have had their impact on the way the data were collected for these statistics.

For a long period, until the year 1895, statistical data collection was based on complete enumeration of populations. The censuses were mostly conducted to establish the size of the population, to determine tax obligations of the people, and to measure the military strength of the country. The idea of sampling had not emerged yet.

The year 1895 marks a fundamental change. Populations had grown bigger and bigger. It was the period of industrialization. Centralized governments required more and more information. The time was ripe for sample surveys. The first ideas emerged around 1895. There was a lot of discussion between 1895 and 1934 about how to select samples: by means of probability sampling or some other sample selection technique.

By 1934, it was clear that only surveys based on probability sampling could provide reliable and accurate estimates. Such methods of data collection were accepted as a scientific. In the period from 1940s to the 1970s, most sample surveys were probability based. Questionnaires were on paper forms. They were completed in face-to-face, telephone, or mail.

Handbook of Web Surveys, Second Edition. Silvia Biffignandi and Jelke Bethlehem.

Somewhere in the 1970s another significant development started. The fast development of microcomputers made it possible to introduce computer-assisted interviewing (CAI). This made survey data collection faster, cheaper, and easier and increased data quality. It was time in which acronyms like CATI (computer-assisted telephone interviewing) and CAPI (computer-assisted personal interviewing) emerged.

The next major development was the creation of the Internet around 1982. When more and more persons and companies got access to the Internet, it became possible to use this network for survey data collection. The first Internet surveys were e-mail surveys. In 1989 the World Wide Web was developed. This software allowed for friendly graphical user interfaces for Internet users. The first browsers emerged and the use of Internet exploded. In the middle of 1990s, the World Wide Web became widely available, and e-mail surveys were increasingly replaced by web surveys.

Web surveys are attractive because they have a number of advantages. They allow for simple, fast, and cheap access to large groups of potential respondents. Not surprisingly, the number of conducted web surveys has increased rapidly over time. There are, however, also potential methodological problems. There are ample examples of web surveys not based on probability sampling. Therefore, generalization of survey results to the population is questionable. The interviewed may access the Internet using various types of devices. Thus, web surveys can be completed and received not only on personal computer (PC) or laptop; it is highly probable the survey to be received in the mobile phone. The so-called mobile web surveys are fully part of web surveys. This implies some methodological problems to be considered, and further research on the impact of mobile is called for.

This chapter describes the historical developments that have led to the emergence of web surveys. As an illustration, Section 1.3 shows how these developments were implemented at Statistics Netherlands and led to new software for survey data collection.

1.2 Theory

1.2.1 THE EVERLASTING DEMAND FOR STATISTICAL INFORMATION

The history of data collection for statistics goes back in time for thousands of years. As far back as Babylonian era, a census of agriculture was carried out. This already took place shortly after the invention of the art of writing. The same thing happened in China. This empire 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 censuses of people and of property. The collected data were used to establish the political status of citizens and to assess their military and tax obligations to the state.

Censuses were rare in the Middle Ages. The most famous one was the census of England taken by the order of William the Conqueror, King of England. The compilation of his *Domesday Book* started in the year 1086 AD. The book records a wealth of information about each manor and each village in the country. Collected information was about more than 13,000 places. More than 10,000 facts were recorded for each country.

To collect all this data, the country was divided into a number of regions. In each region, a group of commissioners was appointed from among the greater lords. Each county within a region was dealt with separately. Sessions were organized in each county town. The commissioners summoned all those required to appear before them. They had prepared a standard list of questions. For example, there were questions about the owner of the manor; the number of free man and slaves; the area of woodland, pasture, and meadow; the number of mills and fishponds, to the total value; and the prospects of getting more profit. The *Domesday Book* still exists, and many county data files are available on CD-ROM and the Internet.

Another interesting example of the history of official statistics is in the Inca Empire that existed between 1000 and 1500 AD. Each Inca tribe had its own statistician, called the *quipucamayoc*. This man kept records of 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 recorded on *quipus*, a system of knots in colored ropes. A decimal system was used for this. At regular intervals, couriers brought the quipus to Cusco, the capital of the kingdom, where all regional statistics were compiled into national statistics. The system of quipucamayocs and quipus worked remarkably well. The system vanished with the fall of the empire.

An early census also took place in Canada in 1666. Jean Talon, the intendant of New France, ordered an official census of the colony to measure the increase in population since the founding of Quebec in 1608. Name, age, sex, marital status, and occupation were recorded for every person. It turned out there lived 3,215 people in New France.

The first censuses in Europe took place in the Nordic countries. The first census in Sweden–Finland took place in 1749. Not everyone welcomed the idea of a census. Particularly religious people believed that people should not be counted. They referred to the census ordered by King David in biblical times, which was interrupted by a terrible plague and never completed. Others said that a population count would reveal the strengths and weaknesses of a country to foreign enemies. Nevertheless, censuses took place in more and more countries. The first census in Denmark–Norway has been in 1769. In 1795, at the time of the Batavian Republic under Napoleon's influence, the first count of the population of the Netherlands took place. The new centralized administration wanted to gather quantitative information to devise a new system of electoral constituencies (see Den Dulk and Van Maarseveen, 1990).

In the period until the late 1880s, there were some applications of *partial investigations*. They were statistical inquiries in which only part of a complete

human population has been interviewed. The way the persons were selected from the population was generally unclear and undocumented.

In the second half of the 19th century, so-called monograph studies became popular. They were based on Quetelet's idea of the average man. According to Quetelet, many physical and moral data have a natural variability. This variability can be described by a normal distribution around a fixed, true value. He assumed the existence of something called the *true value*. Quetelet introduced the concept of *average man* ("l'homme moyenne") as a person of which all characteristics were equal to the true value (see Quetelet, 2010, 2012).

The period of the 18th and 19th centuries is called the era of the Industrial Revolution, too. It led to important changes in society, science, and technology. Among many other things, urbanization started from industrialization and democratization. All these developments created new statistical demands. The foundations for many principles of modern statistics were laid. Several central statistical bureaus, statistical societies, conferences, and journals, were established soon after this period. First ideas about survey sampling emerged in the world of official statistics. If a starting year must be chosen, 1895 would be a good candidate. Anders Kiaer, the founder and first director of Statistics Norway, started in this year a fundamental discussion about the use of sampling methods. This discussion led to the development, acceptance, and application of sampling as a scientific method.

Anders Kiaer (1838–1919) was the founder and advocate of the survey method that is now widely applied in official statistics and social research. With the first publication of his ideas in 1895, he started the process that ended in the development of modern survey sampling theory and methods. This process is described in more detail in Bethlehem (2009).

There have been earlier examples of scientific investigations based on samples, but they were lacking proper scientific foundations. The first known attempt of drawing conclusions about a population using only information about part of it was made by the English merchant John Graunt (1662). He estimated the size of the population of London. Graunt surveyed families in a sample of parishes where the registers were well kept. He found that on average there were three burials per year in 11 families. Assuming this ratio to be more or less constant for all parishes and knowing the total number of burials per year in London to be about 13,000, he concluded that the total number of families was approximately 48,000. Putting the average family size at 8, he estimated the population of London to be 384,000. Since this approach lacked a proper scientific foundation, John Graunt could not say how accurate his estimates were.

More than a century later, the French mathematician Pierre-Simon Laplace realized that it was important to have some indication of the accuracy of his estimate of the French population. Laplace (1812) implemented an approach that was more or less similar to that of John Graunt. He selected 30 departments distributed over the area of France in such a way that all types of climate were represented. Moreover, he selected departments in which accurate population records were kept. Using the central limit theorem, Laplace proved that his estimator had a

normal distribution. Unfortunately, he disregarded the fact that sampling was purposively, and not at random. These problems made application of the central limit theorem at least doubtful.

In 1895 Anders Kiaer (1895, 1997), the founder and first director of Statistics Norway, proposed his *representative method*. It was a partial inquiry in which a large number of persons were questioned. Selection of persons was such that a "miniature" of the population was obtained. Anders Kiaer stressed the importance of *representativity*. He argued 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. Example 1.1 describes the Kiaer's experiment about the representative method.

EXAMPLE 1.1 The representative method of Anders Kiaer

Anders Kiaer applied his representative method in Norway. His idea was to survey the population of Norway by selecting a sample of 120,000 people. Enumerators (hired only for this purpose) visited these people and filled in 120,000 forms. About 80,000 of the forms were collected by the representative method and 40,000 forms by a special (but analogue) method in areas where the working-class people lived.

For the first sample of 80,000 respondents, data from the 1891 census were used to divide the households in Norway into two strata. Approximately 20,000 people were selected from urban areas and the rest from rural areas.

There was a selection of 13 representative cities from the 61 cities in Norway. All five cities having more than 20,000 inhabitants were included, and eight cities representing the medium sized and small towns, too. The proportion of selected people in cities varied: in the middle-sized and small cities, the proportion was greater than in the big cities. Kiaer motivated this choice by the fact that the middle-sized and small cities did not represent only themselves but a larger number of similar cities.

In Kristiania (nowadays Oslo) the proportion was 1/16, in the medium-sized towns the proportion varied between 1/12 and 1/9, and in the small towns it was 1/4 or 1/3 of the population.

Based on the census, it was known how many people lived in each of the 400 streets of Kristiania, the capital of Norway. The sorting of the streets was in four categories according to the number of inhabitants. Then, there was the specification of a selection scheme for each category: the adult population enumeration was in 1 out of 20 for the smallest streets. In the second category, the adult population enumeration was in half of the houses

in 1 out of 10 of streets. In the third category, the enumeration concerned one-fourth of the streets, and the enumeration was every fifth house; and in the last category of the biggest streets, the adult population enumeration was on half of the streets and in 1 out of 10 houses in them.

In selecting the streets their distribution over the city was considered to ensure the largest possible dispersion and the "representative character" of the enumerated areas.

In the medium-sized towns, the sample was selected using the same principles, though in a slightly simplified manner. In the smallest towns, the total adult population in three or four houses was enumerated.

The number of informants in each of the 18 counties in the rural part of Norway was decided considering census data. To obtain representativeness, municipalities in each country, it was used a classification according to their main industry, either as agricultural, forestry, industrial, seafaring, or fishing municipalities. In addition, the geographical distribution was considered.

The total number of the representative municipalities amounted to 109, which is six in each county on average. The total number of municipalities was 498.

The selection of people in a municipality was done in relation to the population in different parishes, and so all different municipalities were covered. The final step was to instruct enumerators to follow a specific path. In addition, instruction to the enumerators was to visit different houses situated close to each other. That is, they were supposed to visit not only middle-class houses but also well-to-do houses, poor-looking houses, and one-person houses.

Kiaer did not explain in his papers how he calculated estimates. The main reason probably was that the representative sample construction was as a miniature of the population. This made computations of estimates trivial: the sample mean is the estimate of the population mean, and the estimate of the population total could be attained simply by multiplying the sample total by the inverse of sampling fraction.

A basic problem of the representative method was that there was no way of establishing the precision of population 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, estimates had an approximately normal distribution. From this moment on, there were two methods of sample selection:

- Kiaer's representative method, based on purposive selection, in which representativity played an essential role and for which no measure of the accuracy of the estimates could be obtained;

- Bowley's approach, based on simple random sampling, for which an indication of the accuracy of estimates could be computed.

Both methods existed side by side until 1934. In that year the Polish scientist Jerzy Neyman published his famous paper (see Neyman, 1934). Neyman developed a new theory based on the concept of the confidence interval. By using random selection instead of purposive selection, there was no need any more to make prior assumptions about the population. The contribution of Neyman was not only that he proposed the confidence interval as an indicator for the precision of estimates. He also conducted an empirical evaluation of Italian census data and proved that the representative method based on purposive sampling was not able to provide satisfactory estimates of population characteristics. He established the superiority of random sampling (also referred to as *probability sampling*) over purposive sampling. Consequently, use of purposive sampling was rejected as a scientific sampling method.

Gradually probability sampling found its way into official statistics. More and more national statistical institutes introduced probability sampling for official statistics. However, the process was slow. For example, a first test of a real sample survey using random selection was carried out by Statistics Netherlands only in 1941 (see CBS, 1948). Using a simple random sample of size 30,000 from the population of 1.75 million taxpayers, it was shown that estimates were accurate.

The history of opinion polls goes back to the 1820s, in which period American newspapers attempted to determine political preference of voters just before the presidential election. These early polls did not pay much attention to sampling. Therefore, it was difficult to establish accuracy of results. Such opinion polls were often called *straw polls*. This expression goes back to rural America. Farmers would throw a handful of straws into the air to see which way the wind was blowing.

It took until the 1920s before more attention was paid to sampling aspects. Lienhard (2003) describes how George Gallup worked out new ways to measure interest in newspaper articles. Gallup used *quota sampling*. The idea was to investigate a group of people that could be considered representative for the population. Hundreds of interviewers across the country visited people. Interviewers were given quota for different groups of respondents. They had to interview so many middle-class urban women, so many lower-class rural men, etc. In total, approximately 3,000 interviews were conducted out for a survey.

Gallup's approach was in great contrast with that of the *Literary Digest* magazine, which was at that time the leading polling organization. This magazine conducted regular "America Speaks" polls. It based its predictions on returned questionnaire forms that were sent to addresses taken from telephone directories books and automobile registration lists. The sample size for these polls was on the order of two million people. So the sample size was much larger than that of Gallup's polls.

The presidential election of 1936 turned out to be decisive for both methods. This is described by Utts (1999). Gallup correctly predicted Franklin Roosevelt to be the new president, whereas *Literary Digest* predicted that Alf Landon would

beat Franklin Roosevelt. The prediction based on the very large sample size turned out to be wrong. The explanation was that the sampling technique of *Literary Digest* did not produce representative samples. In the 1930s, cars and telephones were typically owned by middle- and upper-class people. These people tended to vote Republican, whereas lower-class people were more inclined to vote Democrat. Consequently, Republicans were overrepresented in the *Literary Digest* sample.

As a result of this historic mistake, opinion researchers learned that they should rely on more scientific ways of sample selection. They also learned that the way a sample is selected is more important than the size of the sample.

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

1.2.2 TRADITIONAL DATA COLLECTION

There were three modes of data collection in the early days of survey research: face-to-face interviewing, mail interviewing, and telephone interviewing. Each mode had its advantages and disadvantages.

Face-to-face interviewing was already used for the first censuses. Thus, it is not a surprise it was also used for surveys. Face-to-face interviewing means that interviewers visit the persons selected in the sample. Well-trained interviewers will be successful in persuading reluctant persons to participate in the survey. Therefore, response rates of face-to-face surveys are usually higher than surveys not involving interviewers (for example, mail surveys). Interviewers can also assist respondents in giving the right answers to the questions. This often results in better data. However, the presence of interviewers can also be a drawback. Research suggests that respondents are more inclined to answer sensitive questions properly if there are no interviewers present.

Survey agencies often send a letter announcing the visit of the interviewer. Such a letter can also give additional information about the survey, explain why it is important to participate, and assure that the collected information is treated confidentially. As a result, the respondents are not taken by surprise by the interviewers.

The response rate of a face-to-face survey is usually high and so is quality of the collected data. But a price has to be paid literally: face-to-face interviewing is much more expensive. A team of interviewers has to be trained and paid. They also have to travel, which costs time and money.

Mail interviewing is much less expensive than face-to-face interviewing. Paper questionnaires are sent by mail to persons selected in the sample. They are invited to answer the questions and to return the completed questionnaire to the survey

agency. A mail survey is not interviewers based. Therefore, it is a cheaper mode of data collection than face-to-face survey. Data collection involve mailing costs (letters, postage, envelopes) both for sending the questionnaire and for delivering the questionnaire back; similar costs have to be considered for each reminder. Therefore, costs for stamps and questionnaire printing could be not completely irrelevant. Another advantage of mail survey is that the absence of interviewers can be experienced as less threatening for potential respondents. Therefore, respondents are more inclined to answer sensitive questions properly.

The absence of interviewers also has a number of disadvantages. There are no interviewers to explain questions or assist respondents in answering them. This may cause respondents to misinterpret questions, which has a negative impact on the quality of the collected data. Furthermore, it is not possible to use show cards. A *show card* is typically used for answering closed questions. Such a card contains the list of all possible answers to a question. Respondents can read through the list at their own pace and select the answer corresponding to their situation or opinion. Mail surveys put high demands on the design of the paper questionnaire. For example, it should be clear to all respondents how to navigate through the questionnaire and how to answer questions.

Since the persuasive power of the interviewers is absent, response rates of mail surveys tend to be low. Of course, reminder letters can be sent, but this is often not very successful to let the response rate become very high. More often survey questionnaire forms end up in the pile of old newspapers.

In summary, the costs of a mail survey are relatively low, but often a price has to be paid in terms of data quality: response rates tend to be low, and also the quality of the collected data is also often not very good. Dillman (2007) believes, however, that good results can be obtained by applying his *Tailored Design Method*. This is a set of guidelines for designing and formatting mail survey questionnaires. They pay attention to all aspects of the survey process that may affect response rates or data quality.

Face-to-face interviewing was preferred in the early days of survey interviewing in the Netherlands. The idea was in the 1940s that poor people had poor writing skills and they were not interested in the topics of the surveys. Therefore, they had a lower probability to complete mail questionnaires. People completing and returning questionnaire forms were assumed to be more interested in the survey topics because their intelligence and socioeconomic position was above average.

A third mode of data collection is *telephone interviewing*. Interviewers are needed to conduct a telephone survey, but not as many as for a face-to-face survey, since they do not have to travel from one respondent to the next. They can remain in the call center of the survey agency and conduct more interviews in the same amount of time. Therefore, interviewer costs are less high. An advantage of telephone interviewing over face-to-face interviewing is that respondents may be more inclined to answer sensitive questions, because the interviewer is not present in the room. A drawback in the first days of telephone surveys may be that telephone coverage in the population was low. Not every respondent could be contacted by telephone.

Telephone interviewing has some limitations. Interviews cannot last too long, and question answer may not be written. Obviously, no show cards can be used; lists can be presented by reading them out loud (by the interviewers).

This implies a possible recency effect in the answers. Another problem may be the lack of a proper sampling frame. Telephone directories may suffer from severe under-coverage because many people do not want their phone number to be listed in the directory. Another new development is that increasingly people replace their landline phone by a mobile phone. This fact increases under-coverage in the telephone directories. For example, according to Cobben and Bethlehem (2005), only between 60% and 70% of the Dutch population can be reached through a telephone dictionary. However it has to be advised that mobile phone numbers are not listed in directories in many countries. Thus, a problem arises.

Example 1.2 is about the first telephone survey in the Netherlands.

EXAMPLE 1.2 The first telephone survey in the Netherlands

The first telephone survey took place in the Netherlands on June 11, 1946. See NIPO (1946) for a detailed description. An interview to a few hundred owners of telephones in Amsterdam asked a few questions about listening to the radio. The call to the people was between 20:00 and 21:30 hours on a Tuesday night. Some results are in Table 1.1.

TABLE 1.1 The first telephone survey in the Netherlands

Are you listening to the radio at this moment?	Percentage
Was listening	24
Was not listening	38
Line busy	5
No answer	31
Did not have a radio	2

If people declared they were listening to the radio, the program they were listening to was asked. It turned out that 85% was listening the "Bonte Dinsdagavondtrein," a very famous radio show at that time.

Telephone interviewing has some limitations. Interviews cannot last too long, and no written answer is possible. Obviously, no show cards are used; lists are read loud (by the interviewers).

This implies a possible recency effect in the answers. Another problem may be the lack of a proper sampling frame. Telephone directories may suffer from severe under-coverage because many people do not want their phone number is in the

directory. Another new development is that increasingly people replace their landline phone by a mobile phone. This fact increases under-coverage in the telephone directories. For example, according to Cobben and Bethlehem (2005), only between 60% and 70% of the Dutch population had at that time a telephone dictionary.

A way to avoid the under-coverage problems of telephone directories is to apply *random digit dialing* (RDD) to generate random phone numbers. A computer algorithm computes valid random telephone numbers. Such an algorithm is able to generate both listed and unlisted numbers. Thus, there is complete coverage. An example of an algorithm used in the United Kingdom is to take a number from a directory and replace its last digit by a random digit. RDD also has drawbacks. In some country it is not clear what an unanswered number means. It can mean that the number is not in use. This is a case of over-coverage. No follow-up is needed. It can also mean that someone simply does not answer the phone, which is a case of nonresponse, which has to be followed up. Another drawback of RDD is that there is no information at all about nonrespondents. This makes correction for nonresponse very difficult (Bethlehem, Cobben, and Schouten (2011), see also Chapter 12).

The choice of the mode of data collection is not any easy one. It is usually a compromise between quality and costs. In large countries (like the United States) or sparsely populated countries (like Sweden), it is almost impossible to collect survey data by means of face-to-face interviewing. It requires so many interviewers that have to do so much traveling that the costs would be very high. Therefore, it is not surprising that telephone interviewing emerged here as a major data collection mode. In a very small and densely populated country, like the Netherlands, face-to-face interviewing is much more attractive. Coverage problems of telephone directories and low response rates also play a role in the choice for face-to-face interviewing. More about data collection issues is in Couper et al. (1998).

1.2.3 THE ERA OF COMPUTER-ASSISTED INTERVIEWING

Collecting survey data can be a costly and time-consuming process, particularly if high-quality data are required, the sample is large, and the questionnaire is long and complex. Another problem of traditional data collection is that the completed paper questionnaire forms may contain many errors. Substantial resources must therefore be devoted to cleaning the data. Extensive data editing is required to obtain data of acceptable quality.

Rapid developments in information technology since the 1970s have made it possible to reduce these problems. By introducing microcomputers for data collection, important innovation in surveys took place. A computer program for asking questions and recording the answers replaced the paper questionnaire.

The computer took control of the interviewing process, and it checked answers to the questions. Thus, *computer-assisted interviewing* (CAI) emerged.

CAI comes in different modes of data collection. The first mode of data collection that emerged was *computer-assisted telephone interviewing* (CATI). Couper

and Nicholls (1998) describe its development in the United States in the early 1970s. The first nationwide telephone facility for surveys was established in 1966. The idea at that time was not implementation of CAI but simplifying sample management. The initial systems evolved in subsequent years into full featured CATI systems. Particularly in the United States, there was a rapid growth of the use of these systems. CATI systems were little used in Europe until the early 1980s.

Interviewers in a CATI survey operate a computer running interview software. When instructed to do so by the software, they attempt to contact a selected person by telephone. If this is successful and the person is willing to participate in the survey, the interviewer starts the interviewing program. The first question appears on the screen. If correctly answered, the software proceeds to the next question on the route through the questionnaire.

Call management is an important component of the CATI systems. Its main function is to offer the right telephone number at the right moment to the right interviewer. This is particularly important in cases in which the interviewer has made an appointment with a respondent for a specific time and date. Such a call management system also has facilities to deal with special situations like a busy number (try again after a short while) or no answer (try again later). This all helps to increase the response rate. More about the use of CATI in the United States is in Nicholls and Groves (1986).

Small portable computers came on the market in the 1980s. This made it possible for the interviewers to take computers with them to the respondents. This is the computer-assisted form of face-to-face interviewing, called *computer-assisted personal interviewing* (CAPI). After interviewers have obtained cooperation of the respondents, they start the interviewing program. Questions display is one at a time. Only after the entering of the answer, the next question appeared on the screen.

At first, it was not completely clear whether this mode of data collection could use the computer. There were issues like the weight and size of the computer, the readability of the screen, battery capacity, and the size of keys on the keyboard. Experiments showed that CAPI was feasible. It became clear that CAI for data collection has three major advantages:

- It simplifies the work of interviewers. They do not have to pay attention any more to choosing the correct route through the questionnaire. The computer determines the next question to ask. Interviewers can concentrate more on asking questions and helping respondents giving the proper answers.
- It improves the quality of the collected data. Answers checking is by the software during the interview. Correction of the detected errors is automatic. The respondent is there to provide the proper information. This is much more effective than having to do data editing afterward in the survey agency and without the respondent.
- Data entering in the computer is immediate, during the interview. Straightaway checks are undertaken and detected errors corrected. Therefore, the

record of a respondent is "clean" after completion of the interview. No more subsequent data entry and/or data editing is required. Compared with the old days of traditional data collection with paper forms, this considerably reduces time needed to process the survey data. Therefore, timeliness of the survey results is improved.

Or more information about CAPI in general, see Couper et al. (1998).

The computer-assisted mode of mail interviewing also emerged. It was called *computer-assisted self-interviewing* (CASI), or sometimes also *computer-assisted self-administered questionnaires* (CASAQ). The electronic questionnaire program is sent to the respondents. They run the software, which asks the questions and stores the answers. After the interview completion, the data are send back to the survey agency. Early CASI applications used diskettes or a telephone and modem to transmit the questionnaire and the answers to the question. Later it became common practice to use the Internet as a transport medium.

A CASI survey is only feasible if all respondents have a computer on which they can run the interview program. Since the use of computers was more widespread among companies than among households in the early days of CASI, the first CASI applications were business surveys. An example is the production of fire statistics in the Netherlands in the 1980s. Since all brigades had a microcomputer at that time, data for these statistics CASI were a mode of data collection. Diskettes were sent to the fire brigades. They ran the questionnaire on their MS-DOS computers. The answers were stored on the diskette. After having completed the questionnaire, the diskette was returned to Statistics Netherlands.

An early application in social surveys was the *Telepanel*, set up by Saris (1998). The Telepanel started in 1986. It was a panel of 2,000 households. They agreed on regularly completing questionnaires with the computer equipment provided to them by the survey organization. A home computer was installed in each household. It was connected to the telephone with a modem. It was connected to the television set in the household also. Then it was possible to use it as a monitor. After inserting the diskette into the home computer, it automatically established a connection with the survey agency to exchange information (downloading a new questionnaire or uploading answers of the current questionnaires). Panel members completed a questionnaire each weekend. The Telepanel was in essence very similar to the web panels, which are frequently used nowadays. The only difference was the Internet did not exist yet.

1.2.4 THE CONQUEST OF THE WEB

The development of the Internet started in the early 1970s. The first step was to create networks of computers. The U.S. Department of Defense decided to connect computers of research institutes. Computers were expensive. A network made it possible for these institutes to share each other's computer resources. The name of this first network was ARPANET.

ARPANET became a public network in 1972. Software to send messages over the network was developed. Thus, e-mail was born. Ray Tomlinson of ARPANET was sending the first e-mail in 1971.

The Internet was fairly chaotic in the first decade of its existence. There were many competing techniques and protocols. In 1982, the TCP/IP set of protocols was adopted as the standard for communication of connected networks. This can be seen as the real start of the Internet.

Tim Berners-Lee and scientists at CERN, the European Organization for Nuclear Research in Geneva, were interested in making it easier to retrieve research documentation over the Internet. This led in 1989 to the *hypertext* concept, a text containing references (hyperlinks) to other texts the reader can immediately access. To be able to view these text pages and navigate to other pages through the hyperlinks, Berners-Lee developed a computer software. He called this program a *browser*. The name of the first browser was *World Wide Web*. Now this name denotes the whole set of linked hypertext documents on the Internet.

In 1993, Marc Andreessen and his team at the National Center for Supercomputing Applications (NCSA) (Illinois, USA) developed the browser *Mosaic X*. It was easy to install and use. This browser had increased graphic capabilities. It already contained many features that are common in current browsers. It became a popular browser, which helped to spread the use of the World Wide Web across the world.

The rapid development of the Internet led to new modes of data collection. Already in the 1980s, prior to the widespread introduction of the World Wide Web, e-mail was explored as a new mode of survey data collection. Kiesler and Sproul (1986) describe an early experiment conducted in 1983. They compared an e-mail survey with a traditional mail survey. They showed that the costs of an e-mail survey were much less than those of a mail survey. The response rate of the e-mail survey was 67%, and this was somewhat smaller than the response rate of the mail survey (75%). The turnaround time of the e-mail survey was much shorter. There were less socially desirable answers and less incomplete answers. Kiesler and Sproul (1986) noted that limited Internet coverage restricted wide-scale use of e-mail surveys. In their view, this type of data collection was only useful for communities and organizations with access to and familiarity with computers. These were relatively well-educated, urban, white-collar, and technologically sophisticated people.

Schaefer and Dillman (1998) also compared e-mail surveys with mail surveys. They applied knowledge about mail surveys to e-mail surveys and developed an e-mail survey methodology. They also proposed mixed-mode surveys for populations with limited Internet coverage. They pointed out some advantages of e-mail surveys. In the first place, e-mail surveys could be conducted very fast, even faster than telephone surveys. This was particularly the case for large surveys, where the number of available telephones and interviewers may limit the number of cases that can be completed each day. In the second place, e-mail surveys were inexpensive, because there were no mailing, printing, and interviewer's costs.

The experiment of Schaefer and Dillman (1998) showed that response rates of e-mail and mail surveys were comparable, but the completed questionnaires of the e-mail survey were received much quicker. The answers to open questions were, on average, longer for e-mail surveys. This did not come as a surprise because of the relative ease of typing an answer on a computer compared to writing an answer on paper. There was lower item nonresponse for the e-mail survey. A possible explanation was that moving to a different question in an e-mail survey is much more difficult than moving to a different question on a paper form.

Couper, Blair, and Triplett (1999) found lower response rates for e-mail surveys in an experiment with a survey among employees of statistical agencies in the United States. They pointed out that nonresponse can partly be explained by delivery problems of the e-mails and not by refusal to participate in the survey. For example, if people do not check their e-mail or if the e-mail with the questionnaire does not pass a spam filter, people will not be aware of the invitation to participate in a survey.

Most e-mail surveys could not be seen as a form of CAI. It was merely the electronic analogue of a paper form. There was no automatic routing and no error checking. See Figure 1.1 for a simple example of an e-mail survey questionnaire. It is sent to the respondents. They are asked to reply to the original message. Then they answer the questions in the questionnaire in the reply message. For closed questions they do that by typing an X between the brackets of the option of their choice. The answer to an open question is typed between the corresponding brackets. After completion, they send the e-mail message to the survey agency.

1. What is your age?
   [ ]

2. Are you male or female?
   [ ] Male
   [ ] Female

3. What is your marital status?
   [ ] Married
   [ ] Not married

4. Do you have a job?
   [ ] Yes
   [ ] No

5. What kind of job do you have?
   [write your job]

6. What is your yearly income?
   [ ] Less than 20,000
   [ ] Between 20,000 and 40,000
   [ ] More than 40,000

FIGURE 1.1 **Example of an e-mail survey questionnaire**

Use of e-mail imposes substantial restrictions on the layout. Example 1.3 describes a first way adopted to approach businesses for a web survey. Due to e-mail software of the respondent and the settings of the software, the questionnaire may look different to different respondents. For example, to avoid problems caused by line wrapping, Schaefer and Dillman (1998) advise a line length of at most 70 characters. Schaefer and Dillman (1998) also noted another potential problem of e-mail surveys: the lack of anonymity of e-mail. If respondents reply to the e-mail with the questionnaire, it is difficult to remove all identifying information. Some companies have the possibility to monitor the e-mails of their employees. If this is the case, it may become difficult to obtain high response rates and true answers to the questions asked.

Personalization may help to increase response rates in mail surveys. Therefore, this principle should also be applied to e-mail surveys. An e-mail to a long list of addresses does not help to create the impression of personal treatment. It is probably better to send a separate e-mail to each selected person individually.

EXAMPLE 1.3 The first e-mail survey at Statistics Netherlands

The first test with an e-mail survey at Statistics Netherlands was carried out in 1998. At the time, Internet browsers and HTML were not sufficiently developed and used to make a web survey feasible.

Objective of the test was to explore to what extent e-mail could be used to collect data for the survey on short-term indicators. This was a non-compulsory panel survey, where companies answered a small number of questions about production expectations, order-books, and stocks.

The traditionally mode of data collection for this survey was a mail survey.

The test was conducted in one of the waves of the survey. 1,600 companies were asked to participate in the test. If they did, they had to provide their e-mail address. About 190 companies agreed to participate. These were mainly larger companies with a well-developed computer infrastructure.

A simple text form was sent to these companies by means of e-mail. After activating the reply option, respondents could fill in answers in the text. It was a software-independent and platform-independent solution, but rather primitive from a respondent's point of view.

The test was a success. The response rate among the participating companies was almost 90%. No technical problems were encountered. Overall, respondents were positive. However, they considered the text-based questionnaire old-fashioned, and not very user friendly.

More details about this first test with an e-mail survey at Statistics Netherlands can be found in Roos, Jaspers, and Snijkers (1999).

It should be noted that e-mail can also be used in a different way to send a questionnaire to a respondent. An electronic questionnaire can be offered as an executable file that is attached to the e-mail. The respondents download this interview program on their computers and run it. The advantage of this approach is that such a computer program can have a better graphical user interface. Such a program can also include routing instructions and checks. This way of data collection is sometimes called CASI. Example 1.4 describe an example of a CASI approach.

EXAMPLE 1.4 The production statistics pilot at Statistics Netherlands

In October 2004, Statistics Netherlands started a pilot to find out whether a CASI approach could be used to collect data for yearly production statistics.

One of the approaches tested is denoted by Electronic Data Reporting (EDR). It was a system for responding companies to manage interviewing programs (generated by the Blaise system) on their own computers. The EDR software was sent to respondents on CD-ROM, or respondents could download the software from the Internet.

After the installation of the software, new survey interviews could be sent to respondents by e-mail. These electronic questionnaires were automatically imported in the EDR environment. A simple click would start the interview. After offline completion of the interview, the entered data were automatically encrypted and sent to Statistics Netherlands.

The pilot made clear that downloading the software was feasible. It should be preferred over sending a CD-ROM because it was simpler to manage and less expensive, too. Some companies experienced problems with downloading and installing the software, because security settings of their computer systems and networks prevented them of doing so. User-friendliness and ease of navigation turned out to be important issues for respondents.

For more information about this pilot, see Snijkers, Tonglet, and Onat (2004, 2005).

This form of CASI also has disadvantages. It requires respondents to have computer skills. They should be able to download and run the interviewing program. Couper, Blair, and Triplett (1999) also note that problems may be caused by that fact that different users may have different operating systems on their computers or different versions of the same operating system. This may require different versions of the interviewing program, and it must be known in advance which operating system a respondent has. Moreover, the size of an executable file may be substantial, which may complicate sending it by e-mail.

E-mail surveys had the advantages of speed and low costs. Compared with CAI they had the disadvantages of a poor user interface and lack of adequate

editing and navigation facilities. An e-mail questionnaire was just a paper questionnaire in an e-mail. The Internet became more interesting for survey data collection after HTML 2.0 was introduced in 1995. HTML stands for Hypertext Markup Language. It is the markup language for web pages. The first version of HTML was developed by Tim Berners-Lee in 1991. Version 2 of HTML included support for forms. This made it possible to transfer data from a user to the web server. Web pages could contain questions, and the answers could be collected by the server. Example 1.5 shows some applicative aspects of the HTML questions.

EXAMPLE 1.5 Designing questions in HTML 2.0

Version 2.0 of HTML made it possible to implement questions on a web page. The `<input>` tag can be used to define different types of questions. With `type=radio` this tag becomes a *radio button*. A *closed question* is defined by introducing a radio button of each possible answer. See Figure 1.2 for an example. Not more than one radio button can be selected. This corresponds to a closed question for which only one answer must be selected.

FIGURE 1.2 **A closed question in HTML**

Sometimes respondents must be offered the possibility to select more than one answer, like in Figure 1.3. Respondents are asked for their means of transport to work. Some people may use several transport means.

FIGURE 1.3 **A check-all-that-apply question in HTML**

For example, a person may first take a bicycle to the railway station and then continues by train. Such a closed question is sometimes also called a *check-all-that-apply* question. It can be implemented in HTML by means of a

series of *checkboxes*. A checkbox is obtained by stetting the type of the `<input>` tag to `checkbox`.

Figure 1.4 shows the implementation of an open question. Any text can be entered in the input field. A limit may be set to the length of the text. An open question is defined with `type=text` for the `<input>` tag.

FIGURE 1.4 **An open question in HTML**

If an input field is preferred that allows for more lines of text to be answered, the `<textarea>` tag can be used for this.

There are no specific types of the `<input>` tag for other types of questions. However, most of these question types can be implemented with the input field of an open question. For example, Figure 1.5 shows a numeric question. The question is basically an open question, but extra checks on the answer only allow numbers to be entered within certain bounds.

FIGURE 1.5 **A numeric question in HTML**

Date question can be specified as a set of three input fields: one for the day, one for the month, and one for the year.

In the first years of the World Wide Web, use of web surveys was limited by the low penetration of the Internet. Internet penetration was higher among establishments than among households. Therefore, it is not surprising that first experiments tested the use of web business surveys. Clayton and Werking (1998) describe a pilot carried out in 1996 for Current Employment Statistics (CES) program of the U.S. Bureau of Labor Statistics. They expected the web to offer a low-cost survey environment. Because it was a form of true online data collection, an immediate response to the answers of the respondents was possible. This could improve data quality. They also saw the great flexibility of web survey

questionnaires. They could be offered in a form layout or in a question-by-question approach. The drawback was the limited number of respondents having access to the Internet. Only 11% of CES respondents had access to Internet and a compatible browser.

Roos and Wings (2000) conducted a test with Internet data collection at Statistics Netherlands for the construction industry. Respondents could choose between three modes:

- Completing a form offline. The form was sent as an HTML file that was attached to an e-mail. The form is downloaded, completed offline, and returned by e-mail.
- Completing a form online. The Internet address of an online web form was sent by e-mail. The form was completed online.
- Completing an e-mail form. An e-mail is sent containing the questionnaire in plain text. Respondents clicked the reply button, answered the questions, and sent the e-mail back.

A sample of 1,500 companies was invited to participate in the experiment. 188 companies were willing and able to participate. Of those, 149 could surf the Internet, and 39 only had e-mail. Questionnaire completion times of all three modes were similar to that of a paper form. Respondents preferred the form-based layout over the question-by-question layout. The conclusion of the experiment was that web surveys worked well.

General population web surveys were rare in the first period of existence of the Internet. This was due to the low Internet penetration among households. This prevented conducting representative surveys. However, there were polls on the Internet. Recruitment of respondents was based on self-selection and not on probability sampling. Users could even create their own polls on websites like *Survey Central*, *Open Debate,* and *Internet Voice* (see O'Connell, 1998).

Also in 1998, the *Survey2000* project was carried out. This was a large self-selection web survey on the website of the National Geographic Society. This was a survey on mobility, community, and cultural identity. In a period of two months, over 80,000 respondents completed the questionnaire. See Witte, Amoroso, and Howard (2000) for more details about this project.

It seems to be typical for this type of self-selection web surveys that they make it possible to collect data about a large number of respondents in a relatively short time. Other examples are given by Bethlehem and Stoop (2007). The survey *21minuten.nl* has been conducted a number of times in the Netherlands. This survey supposed to supply answers to questions about important problems in Dutch society. Within a period of six weeks in 2006, about 170,000 people completed the online questionnaires. A similar survey was conducted in Germany. It is called *Perspektive Deutschland*. More than 600,000 participated in this survey in 2005/2006.

It should be noted that these large sample sizes are no guarantee for proper statistical inference. Due to under-coverage (not everyone has access to the

Internet) and self-selection (no proper random sampling), estimates can be biased. This bias is independent of the sample size.

Internet penetration is still low in many countries, making it almost impossible to conduct a general population web survey. Since data collection costs can be reduced if the Internet is used, other approaches are sought. One such approach is *mixed-mode data collection*. A web survey is combined with one or more other modes of data collection, like a mail survey, a telephone survey, or a face-to-face survey. Researchers first attempt to collect as much data as possible with the cheapest mode of data collection (web). Then, the nonrespondents are re-approached in a different (next cheapest) mode. Example 1.6 describes a survey run using a mixed-mode approach.

EXAMPLE 1.6 Experiment with a mixed-mode surveys

Beukenhorst and Wetzels (2009) describe a mixed-mode experiment conducted by Statistics Netherlands. They used the Dutch Safety Monitor for this experiment. This survey asks questions about feelings of security, quality of life, and level of crime experienced. The sample for this survey was selected from the Dutch population register. All sampled persons received a letter in which they were asked to complete the survey questionnaire on the Internet. The letter also included a postcard that could be used to request a paper questionnaire. Two reminders were sent to those that did not respond by web or mail. If still no response was obtained, nonrespondents were approached by means of CATI, if a listed telephone number was available. If not, these nonrespondents were approached by CAPI.

To be able to compare this four-mode survey with a traditional survey, also a two-mode survey was conducted for an independent sample. Sampled persons were approached by CATI if their telephone number was listed in the directory, and otherwise they were approached by CAPI.

The response rate for four-mode survey turned out to be 59.7%. The response rate for the two-mode survey was higher. So, introducing more modes did not increase the overall response rate. However, more than half of the response (58%) in the four-mode survey was obtained with a self-administered mode of data collection (web or paper). Therefore, the costs of the survey were much lower. Interviewers were deployed in only 42% of the cases. For more detail, see Beukenhorst and Wetzels (2009) or Bethlehem, Cobben, and Schouten (2011).

A special case of mixed-mode data collection is related to the increasing diffusion of mobile phones and smartphones. When an invitation e-mail is sent, the questionnaire might be received either on a computer or a mobile phone or a

smartphone. The interviewee could complete the web questionnaire using either the mobile device or the computer. Thus, it is better to talk about mobile web surveys rather than web survey. A recommendation is to run web surveys that are fully adapt for smartphones. Therefore, in presenting methods for web surveys, comments about the adaption for smartphone surveys will be discussed all along the chapters of this handbook. Some penetration data allow for understanding how the situation differs across the countries. The coverage of telephone directories, of Internet, and of mobile cells provides the feeling of the need to adopt a mixed-mode approach. A World Bank study reports that, in the 2018, Euro area fixed telephone subscription for 100 people is 44.4, mobile 122.6 and Internet 83.8. Table 1.2 shows the same indicators by country. Only some countries are shown in the table since the objective is just to evidence that there is a relevant difference across countries.

TABLE 1.2 Penetration of fixed and mobile phone and of Internet (year 2018)

Country	Fixed telephone subscription (% of inhabitants)	Mobile cellular subscription (% of inhabitants)	% of individuals using the Internet
Austria	42	125	88.0
Denmark	19	125	97.6
Finland	6	132	88.9
France	59	108	82.0
Germany	52	129	89.7
Greece	47	116	72.9
Italy	34	137	74.4
The Netherlands	35	124	94.7
Norway	11	107	96.5
Portugal	50	115	74.7
Romania	19	116	70.7
Slovenia	33	118	79.7
Spain	40	116	86.1
Sweden	24	125	92.1
Switzerland	39	130	90.0

Source: Data from International Telecommunication Union. World Telecommunication/ICT.

Note that Internet users are individuals who have used the Internet (from any location) in the last 12 months. Internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV, etc. Fixed telephone subscriptions refer to the sum of active number of analogue fixed telephone lines, voice-over-IP (VoIP) subscriptions, fixed wireless local loop (WLL) subscriptions, ISDN voice-channel equivalents, and fixed public payphones. Mobile cellular telephone subscriptions are subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology.

1.2.5 WEB SURVEYS AND OTHER SOURCES

Current digital environment and technology trends are providing a huge amount of data about most phenomena. These data are available on the web. Often they are free of charge, if not protected for privacy.

Examples of data available in digital format are credit card transactions, tax data, social chatting, telephone use (calls details: time, location, length of the call, etc.), social security payments, GPS, videos. Using this type of data for statistical purposes is appealing and challenging. The term big data is currently used to characterize data with high volume, velocity, and variety. There is a debate on the definition and on the use of big data for statistical purposes.

Roughly speaking, big data are based on the automatic collection on everything that people do; they are not subject to statistical classification criteria and to statistical treatment for representativity. Also administrative data, i.e., information that are collected for registering units (people, businesses, sales, and so on) into an activity process, might be included into big data.

Practitioners and researchers are now wondering if big data could substitute web surveys to provide information for social and economic decision making. For a discussion about that, see Couper (2013).

It should be emphasized that conclusion is that big data and web surveys are complementary data sources, not competing data sources. The availability of big data to support research provides a new way to approach old questions as well as an ability to address some new questions that in the past were not considered.

Web surveys may be run as stand-alone surveys. However, source integration is a major trend for the future of the next 10 years of web surveys.

It is a new area of research to achieve the three following goals: (1) Minimize the cost associated with surveys. (2) Maximize the information, i.e., the findings based on big data generate more questions, and some of those questions could be best addressed by web surveys or other traditional survey methods. Moreover, information from one source could be useful for improving data to be estimated from a survey. (3) Minimize the respondent burden. Integration alleviates the burden of duplicating data gathering efforts and enables the extraction of information that would otherwise be impossible.

Therefore, it is necessary to work in the direction of using big data and integrating them with the survey results. It is important to face experimental applications having in mind the characteristics, nature, and the limitations of big data as statistical sources and the methodological soundness of the survey results.

At the time being, market research and private/public businesses have great interest in trying to use big data to investigate markets and individual behavior. The use of this data as exploratory source is the most plausible application, whereas using this data for statistical purposes and integration with web survey requires still a lot of effort around definitions, classifications, and estimation methodological problems.

Official statistics producers are investigating how to use other big data sources and how to produce estimates in a multisource framework. Some experiments have

been already undertaken, with contrasting results. Most successful applications consist in the integration of web survey data and administrative data (i.e., administrative data could be considered a type of big data according to many authors). Administrative data have been used:

- To generate a survey frame or to supplement/update an existing frame. When surveys are run on the web, administrative data integration could help in applying the adaptive survey design (see Chapter 8) to improve the data collection process. An ultimate task could be the replacement of data collection (e.g., use of taxation data for small businesses instead of seeking survey data for them). In this case, however, data replacement would be about some basic data; surveys will be anyway useful for collecting data about specific topics and behaviors. For example, think of all the surveys with a focus on consumption. If big data assets are providing insights on consumption via passive observation, primary research via surveys will not have to collect this type of information, and it is finally possible to deliver on the vision of shorter surveys instead of simply providing complementary data to the desired information. Surveys can be short and focused on those variables that they are ideally suited for, resulting in better data quality.
- In editing and imputation.
- In estimation (e.g., as auxiliary information in calibration estimation, benchmarking, or calendarization).
- In comparing survey estimates with estimates from a related administrative program as well as other forms of survey evaluation have been experienced.

When using a multisource approach in web surveys, several aspects should be considered.

First of all is the heterogeneous nature of the sources with respect to the following basic characteristics: the aggregation level, the unit, the variables, the coverage, the time, the population, and the data type:

The aggregation level, i.e., some data sources consist of only microdata, some other data sources consist of a mix of microdata and aggregated data, whereas in some other cases data sources consist of only aggregated data. In some case, aggregated data are available besides microdata. There is still overlap between the sources, from which there arises the need to reconcile the statistics at some aggregated level. Of particular interest is when the aggregated data are estimates themselves. Otherwise, the conciliation can be achieved by means of calibration, a standard approach in survey sampling.

As regards the units, it has to be considered that sometimes there are no overlapping units in the data sources or only some units are overlapping. Also, as regards the variables, no overlapping variables in the data sources could occur, or only variables in the data sources could overlap.

Under-coverage versus there is no under-coverage has to be considered. The data sources are cross-sectional versus other data sources are longitudinal; thus, the researcher should take care of what type of data he is integrating. The set of

population units from a population register could be known, or the population list is not known; this affects the possibility of generating a probability-based sample. In some cases, a data source contains a complete enumeration of its target population, or a data source is selected by means of probability sampling from its target population, or a data source is selected by non-probability sampling from its population. The database may be further split into two subcases depending on whether one of the data sources consists of sample data (and where the sampling aspects play an important role in the estimation process) or not. In the former case, specific methods should be used in the estimation process, for instance, taking the sampling weights into account and considering that sample data may include specific information that is not reported in register.

Another aspect is the configuration of the sources to be integrated. There are a few basic ways, most commonly encountered. However, in practice, a given situation may well involve several basic configurations at the same time.

The first and most basic configuration of the integration process of different sources is multiple cross-sectional data that together provide a complete data set with full coverage of the target population. Provided they are in an ideal error-free state, the different data sets, or data sources, are complementary to each other and can be simply "added" to each other in order to produce output statistics.

A second type of configuration is when there exists *overlap* between the different data sources. The overlap can concern the units, the measured variables, or both.

A third situation is when the combined data *entail under-coverage of the* target population in addition, even when the data are in an ideal error-free state.

A further configuration is when microdata and aggregated data are available. There is overlap between the sources, but there is the need to reconcile the statistics at some aggregated level. The conciliation can be achieved by means of calibration, which is a standard approach in survey sampling. Of particular interest is when the aggregated data are estimates themselves.

Finally, it is possible that multisource approach refers to longitudinal data. More questions arise; the most important issue is that of reconciling time series of different frequencies and qualities. For example, one source has monthly data and the other source has quarterly data.

Integration may occur between different types of sources: surveys (mainly web surveys), administrative data, other passive collected data, social network, and other unstructured data.

Integration of different configurations as well as of different types of data sources implies different methodological problems. For instance, integrating survey and administrative data through unit record linkage requires improving coherence across data collections, using standard classifications and questions, rationalizing content between surveys, and processes for combining separate sample surveys into one survey vehicle (Bycroft, 2010).

Example 1.7 discusses an integration of web scraped information, administrative data, and surveys, whereas Example 1.8 shows an application of integration between survey data and social network unstructured information.

As a result of the multisource integration, statistical output is based on complex combinations of sources. Its quality depends on the quality of the primary sources and the ways they are combined. Some studies are investigating the appropriateness of the current set of quality measures for multiple source statistics; they explain the need for improvement and outline directions for further work.

EXAMPLE 1.7 Web scraping, administrative data, and surveys

Istat since 2015 has been experimenting web scraping, text mining, and machine learning techniques in order to obtain a subset of the estimates currently produced by the sampling survey on "Survey on ICT Usage and e-Commerce in Enterprises" yearly carried out on the web. Studies from Barcaroli et al., (2015, 2016), and Righi, Barcaroli, and Golini (2017) have focused in implementing the experiment and in evaluating data quality.

Trying to make the optimal use of all available information from the administrative sources to web scraping information, web survey estimates produced could tentatively be improved. The aim of the experiment is also to evaluate the possibility to use the sample of surveyed data as a training set in order to fit models to be applied to website information.

Recent and in progress steps are a further improvement to the performance of the models by adding explicative variables consisting not only of single terms, but the joint consideration of sequences of terms relevant for each characteristic object of interest. When a certain degree of quality of the resulting predictive models will be guaranteed, their application to the whole population of enterprises owning a website will be performed. A crucial task will be also the retrieval of the URLs related to the websites for the whole population of enterprises. Finally, once having predicted the values of the target variables for all reachable units in the population, the quality of estimates obtained will be analyzed and compared with the current sampling estimates obtained by the survey. In a simulation study, Righi, Barcaroli, and Golini (2017) found that the use of auxiliary variable coming from the Internet DB source highly correlated with the target variable does not guarantee enhancement of the quality of the estimates if selectivity affects the source. Bias may occur due to absence of some subgroups. Thus an analysis of the DB variable and the study of the relationship between populations covered or not by the DB source is a fundamental step to know how to use and which framework implement to assure high-quality output.

In conclusion, the approach that uses web scraping and administrative data together with the web survey looks to be promising; nevertheless quality results of the estimations are satisfactory only in some cases. The use of big data has to be carefully evaluated, especially if selectivity affects the source.

Example 1.8 focuses on an experiment of integration of social media data and surveys. Even if the study lacks statistical representativeness and indicators, it presents an interesting approach that should be deeply investigated and statistically formalized.

EXAMPLE 1.8 Social media and surveys

Wells and Thorson (2015) introduce a novel method that combines a "big data" measurement of the content of individuals' Facebook (FB) news feeds with traditional survey measures to explore the antecedents and effects of exposure to news and politics content on the site. This hybrid approach is used to untangle distinct channels of public affairs content within respondents' FB news feeds.

The authors explore why respondents vary in the extent to which they encounter public affairs content on the website. Moreover, they examine whether the amount and type of public affairs content flows in one's FB are associated with political knowledge and participation above and beyond self-report measures of news media use.

To combine a survey with measurements of respondents' actual FB experiences, they created a FB application ("app") and embedded it within an online survey experience.

Respondents, undergraduates at a large Midwestern public university, visited a web page and gave two sets of permissions: they first consented to be participants in a research study—a form required by the institutional review board—and then they separately approved the app through their FB profile. Once they approved the app, they were returned to the survey to complete the questionnaire. While respondents completed the questionnaire, the app recorded specific elements of their FB experience (with respondents' permission), such as how many friends they had, what pages they followed, and what content appeared in their news feeds during the previous week. When respondents had completed the survey, the app had finished its work and automatically removed itself from respondents' profiles. This research was approved by a standard university institutional research board and was designed to comply with FB's Platform Policies and Statement of Rights and Responsibilities, each of which placed restrictions on the use and presentation of the data.

The resulting database offers an original combination of respondent's self-reported attitudes and media behaviors (including FB experience) with measure of part of their FB experience.

From the statistical point of view, the study has limitations (Beręsewicz et al., 2018; Biffignandi and Signorelli, 2016). The empirical study is run on a small sample of college volunteers. Thus, they have no claim of

representativeness. In addition they have considered only a single information platform (FB). Other limitations suggest to consider the results just as a first experimental research. However, the approach proposed is in line with interesting methodological innovations toward the combination of social media trace with conventional methods. It opens the perspective to better understand big data and then try to relate big data descriptive information to socioeconomic theoretical hypotheses.

Obviously, it is underlined that the statistical perspective of representativeness of the results should be considered in future studies. No probability-based sample ad coverage limitations (partial coverage and possibility of duplications) mine to the generalization of the results. New methodological solutions need to be adopted for representativeness of these interesting preliminary results.

1.2.6 HISTORIC SUMMARY

The history above shows that technology changes have impacted survey taking and methods:

- Paper questionnaires were exclusively used for decades until the 1970s and 1980s for both self-completion and by interviewers. Processing the data was expensive and focused on eliminating survey-taking mistakes.
- Computer questionnaires at first were used solely for interviewing, while paper questionnaires were still used for self-completion.
- The advent of the Internet meant that self-completion could now be computer based, but this was limited at first to browsers on PC.
- Computing advances in hardware, software, and connectivity enabled and forced changes in survey taking, processing, and methods.

1.2.7 PRESENT-DAY CHALLENGES AND OPPORTUNITIES

In the past 15 years, rapid technical and social changes have introduced a number of challenges and opportunities. The following is a high-level list of challenges:

- The respondent is much more in charge of the survey including whether and how he/she will participate.
- There is such a vast proliferation of computing devices and platforms that survey takers cannot design and test for each possible platform.
- Modern-day surveys must be accessible to all self-respondents, including the blind, visually impaired, and the motor impaired.

- Few survey practitioners have all the skills needed to effectively design surveys for all platforms and to make them accessible at the same time.

Pierzchala (2016) listed a number of technical challenges that face survey practitioners. This list was developed to communicate the magnitude of the challenges. The term *multis* refers to the multiple ways that surveys may have to adapt for a particular study:

- **Multicultural surveys**: There are differences in respondent understanding, values, and scale spacing due to various cultural norms. These can lead to different question formulation or response patterns.
- **Multi-device surveys**: There are differences in questionnaire appearance and function on desktops, laptops, tablets, and smartphones.
- **Multilingual surveys**: There are translations, system texts, alphabetic versus Asian scripts, left-to-right versus right-to-left scripts, and switching languages in the middle of the survey.
- **Multimode surveys**: There are interviewer- and self-administered surveys such as CATI and CAPI for interviewers and browser and paper self-completion modes (Pierzchala, 2006).
- **Multinational surveys**: There are differences in currency, flags and other images, names of institutions, links, differences in social programs, and data formats such as date display.
- **Multi-operable surveys**: These are differences in how the user interacts with the software and device including touch and gestures versus keyboards with function keys. Whether there is a physical keyboard or a virtual keyboard impacts screen space for question display.
- **Multi-platform surveys**: These are differences in computer operating systems, whether the user is connected or disconnected to/from the server, and settings such as for pop-up blockers.
- **Multi-structural surveys**: There can be differences in question structures due to visual versus aural presentation, memory demands on the respondent, and linear versus nonlinear cognitive processing.
- **Multi-version surveys**: In economic surveys, questionnaires can vary between industries. For example, an agricultural survey asks about different crops in different parts of the country, and different crops can have different questions.

These *multis* lead to changes in question wording, text-presentation standards, interviewer or respondent instructions, location of page breaks, number of questions on a page, question format, allowed responses, whether choices for *don't know* (DK) or *refusal* (R) are explicitly presented or are implied, and whether the user can advance without some kind of answer (even if *DK* or *RF*) or can just proceed at will to the next question or page.

There can be additional challenges. Governmental and scientific surveys can be long and complex. Surveys must be accessible and usable to the disabled.

Additionally, there are ever-tightening constraints including not enough time, not enough people or money, unclear and late and inconsistent specifications, last-minute changes, screens that are too small, and computers that are too slow.

1.2.8 CONCLUSIONS FROM MODERN-DAY CHALLENGES

The description of modern-day survey challenges leads to some conclusions:

- Modern-day surveys can be very hard.
- No single person has all the answers.
- New survey-producing methods are necessary to address all the challenges within ever-tightening constraints.
- Small screen sizes often lead to adaptations of survey instruments such as using fewer points in a scale question.
- With the proliferation of devices, it becomes harder to rely on *unimode* designs where all questions appear the same in all modes and devices (Dillman, Smyth, and Christian, 2014). Instead, the institute may strive for *cognitive equivalence* across all manifestations (de Leeuw, 2005).

1.2.9 THRIVING IN THE MODERN-DAY SURVEY WORLD

Updated survey design methods may give ways to handle and even thrive in the modern-day survey world. The idea is to use extremely powerful computer-based specification to replace document specification and manual programming. This idea is described in the following:

- Use a capable computer-based specification system to define the questionnaire. A drag-and-drop specification may be adequate for simpler surveys, but when you get to surveys that must handle more of the *multis* mentioned above, or when you get to thousands of questions, drag-and-drop becomes too onerous.
- Specification and survey methods research should use question structures (see below).
- The institute should define its question-presentation standards for each structure across all the *multis*. This requires some up-front work and decisions.
- When the specification is entered, the computer should generate the necessary source code and related configuration files for all *multis*. All these computer-generated outputs should conform to the institute's standards.
- Use a survey-taking system that has evolved to cope with the modern-day world.

1.3 Application

1.3.1 BLAISE

The historic developments with respect to surveys as described in the previous section took also place in the Netherlands. Particularly the rapid developments in computer technology have had a major impact on the way Statistics Netherlands collected its data. Efforts to improve the collection and processing of survey data in terms of costs, timeliness, and quality have led to a powerful software system called Blaise. This system emerged in the 1980s, and it has evolved over time so that it is now also able to conduct web surveys and mixed-mode surveys. The section gives an overview of the developments at Statistics Netherlands leading to Internet version of Blaise.

The advance of computer technology since the late 1940s led to many improvements at Statistics Netherlands for conducting its surveys. For example, from 1947 Statistics Netherlands started using probability samples to replace its complete enumerations for surveys on income statistics and agriculture. The implementation of sophisticated sampling techniques such as stratification and systematic sampling is much easier and less labor intensive on a computer than manual methods.

Collecting and processing statistical data was a time-consuming and expensive process. Data editing was an important component of this work. The aim of these data editing activities was to detect and correct errors in the individual records, questionnaires, or forms. This should improve the quality of the results of surveys. Since statistical offices attached much importance to this aspect of the survey process, a large part of human and computer resources were spent on it.

To obtain more insight into the effectiveness of data editing, Statistics Netherlands carried out a Data Editing Research Project in 1984. Bethlehem (1987) describes how survey data were processed. The overall process included manual inspection of paper forms, preparation of the forms for high-speed data entry including correcting obvious errors or following up with respondents, data entry, and further correction.

The Data Editing Research Project discovered a number of problems:

- Various people from different departments were involved. Many people dealt with the information: respondents, subject-matter specialists, data typists, and computer programmers.
- Transfer of material from one person/department to another could be a source of error, misunderstanding, and delay.
- Different computer systems were involved from mainframe to minicomputers to desktop computers under MS-DOS. Transfer of files from one system to another caused delay, and incorrect specification and documentation could produce errors.

- Not all activities were aimed at quality improvement. Time was also spent on just preparing forms for data entry, and not on correcting errors.
- The cycle of data entry, automatic checking, and manual correction was in many cases repeated three times or more. Due to these cycles, data processing was very time consuming.
- The structure of the data (the metadata) had to be specified in nearly every step of the data editing process. Although essentially the same, the "language" of this metadata specification could be completely different for every department or computer system involved.

The conclusions of the Data Editing Research Project led to general redesign of the survey processes of Statistics Netherlands. The idea was to improve the handling of paper questionnaire forms by integrating data entry and data editing tasks. The traditional batch-oriented data editing activities, in which the complete data set was processed as a whole, were replaced by a record-oriented process in which each record (form) was completely dealt with in one session.

More about the development of the Blaise system and its underlying philosophy can be found in Bethlehem and Hofman (2006).

The new group of activities was implemented in a so-called CADI system. CADI stands for *computer-assisted data input*. The CADI system was designed for use by the workers in the subject-matter departments. Data could be processed in two ways by this system:

- *Heads-up data entry*. Subject-matter employees worked through a pile of forms with a microcomputer, processing the forms one by one. First, they entered all data on a form, and then they activated the check option to test for all kinds of errors. Detected errors were reported on the screen. Errors could be corrected by consulting forms or by contacting the suppliers of the information. After elimination of all errors, a "clean" record was written to file. If employees could not produce a clean record, they could write the record to a separate file of "dirty" records to deal with later.
- *Heads-down data entry*. Data typists used the CADI system to enter data beforehand without much error checking. After completion, the CADI system checked in a batch run all records and flagged the incorrect ones. Then subject-matter specialists handled these dirty records one by one and correct the detected errors.

To be able to introduce CADI on a wide scale in the organization, a new standard package called Blaise was developed in 1986. The basis of the system was the Blaise language, which was used to create a formal specification of the structure and contents of the questionnaire.

The first version of the Blaise system ran on networks of microcomputers under MS-DOS. It was intended for use by the people of the subject-matter departments; therefore no computer expert knowledge was needed to use the Blaise system.

In the Blaise philosophy, the first step in carrying out a survey was to design a questionnaire in the Blaise language. Such a specification of the questionnaire contains more information than a traditional paper questionnaire. It did not only describe questions, possible answers, and conditions on the route through the questionnaire but also relationships between answers that had to be checked.

Figure 1.6 contains an example of a simple paper questionnaire. The questionnaire contains one route instruction: persons without job are instructed to skip the questions about the type of job and income.

Figure 1.7 contains the specification of this questionnaire in the Blaise system. The first part of the questionnaire specification is the *Fields section*. It contains the definition of all questions that can be asked. A question consists of an identifying name, the text of the question as presented to the respondents, and a specification of valid answers. For example, the question about age has the name *Age*, the text of the question is "*What is your age?*" and the answer must be a number between 0 and 99.

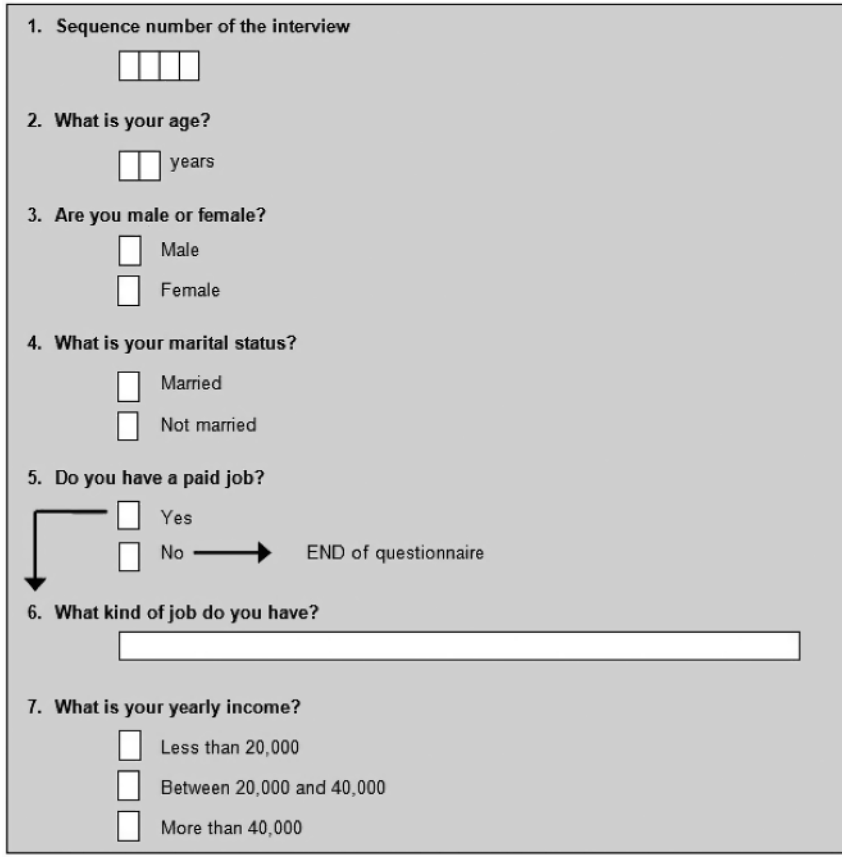

FIGURE 1.6 **A simple paper questionnaire**

```
DATAMODEL LFS "The Labour Force Survey";

FIELDS
    SeqNum  "Sequence number of the interview?": 1..1000
    Age     "What is your age?": 0..99
    Sex     "Are you male or female?": (Male, Female)
    MarStat "What is your marital status?":
               (Married "Married",
                NotMar  "Not married")
    Job      "Do you have a job?": (Yes, No)
    JobDes   "What kind of job do you have?": STRING[20]
    Income   "What is your yearly income?":
               (Less20  "Less than 20,000",
                Upto40  "Between 20,000 and 40,000",
                More40  "More than 40,000")

 RULES
    SeqNum Age Sex MarStat Job
    IF Job = Yes THEN
       JobDes Income
    ENDIF

    IF Age < 15 "respondent is younger than 15" THEN
       MarStat = NotMar "he/she is too young to be married!"
    ENDIF

 ENDMODEL
```

FIGURE 1.7 **A simple Blaise questionnaire specification**

The question *JobDes* requires a text not exceeding 20 characters. *Income* is a closed question. There are three possible answer options. Each option has a name (for example, *Less20*) and a text for the respondent (for example, "*Less than 20,000*").

The second part of the Blaise specification is the *Rules section*. Here, the order of the questions is specified and the conditions under which they are asked. According to the rules section in Figure 1.7, every respondent must answer the questions *SeqNum*, *Age*, *Sex*, *MarStat,* and *Job* in this order. Only persons with a job (*Job = Yes*) have to answer the questions *JobDes* and *Income*.

The rules section can also contain checks on the answers of the questions. Figure 1.7 contains such a check. If people are younger than 15 years (*Age* < 15), then their marital status can only be not married (*MarStat = NotMar*). The check also contains texts that are used to display the error message on the screen (*If respondent is younger than 15 then he/she is too young to be married!*).

The rules section may also contain computations. Such computations could be necessary in complex routing instructions or checks or to derive new variables.

The first version of Blaise used the questionnaire specification to generate a CADI program. Figure 1.8 shows what the computer screen of this MS-DOS program looked like for the Blaise questionnaire in Figure 1.7.

FIGURE 1.8 **A Blaise CADI program**

Since this program was used by subject-matter specialists, only question names are shown on the screen shown in Figure 1.8. Additional information could be displayed through special keys. Note that the input fields for the questions *Age* and *MarStat* contain error counters. These error indicators appeared because the answers of the questions *Age* (2) and *MarStat* (*Married*) did not pass the check.

After Blaise had been in use for a while, it was realized that such a system could be made much more powerful. The questionnaire specification in the Blaise system contained all knowledge about the questionnaire and the data needed for survey processing. Therefore, Blaise should be capable to handle CAI.

Implementing CAI means that the paper questionnaire is replaced by a computer program containing the questions to be asked. The computer takes control of the interviewing process. It performs two important activities:

- *Route control*. The computer program determines which question is to be asked next and displays that question on the screen. Such a decision may depend on the answers to previous questions. As a result, it is not possible anymore to make route errors.
- *Error checking*. The computer program checks the answers as data are entered. Range checks are carried out immediately, as well as consistency checks after entry of all relevant answers. If an error is detected, the program produces an error message, and data must be corrected.

Use of computer-assisted data collection has three major advantages. First, it simplifies the work of interviewer (for example, no more route control). Second, it improves the quality of the collected data. Third, data are entered in the computer during the interview resulting in a complete and clean record.

FIGURE 1.9 **A Blaise CAPI program**

Version 2 of Blaise was completed in 1988. It implemented CAPI. This is a form of face-to-face interviewing in which interviewers use a laptop computer to conduct the interview.

Figure 1.9 shows an example of a screen of a CAPI program generated by Blaise. The screen was divided in two parts. The upper part contains the current question to be answered (*What kind of a job do you have?*). After an answer had been entered, this question was replaced by the next question on the route.

Just displaying one question at the time gave the interviewers only limited feedback on where they are in the questionnaire. Therefore, the lower part of the screen displayed (in a very compact way) the current page of the questionnaire.

Statistics Netherlands started full-scale use of CAPI in regular survey in 1987. The first CAPI survey was the Labor Force Survey. Each month, about 400 interviewers equipped with laptops visited 12,000 addresses. After a day of interviewing, the laptop was connected to a telephone modem. The data were transmitted to the office at night. In return, new addresses were sent to the interviewers. The next morning the laptop was prepared for a new day of interviewing.

CATI was introduced in 1990 on desktop computers. Interviewers called respondents from a central unit (call center) and conducted interviews by telephone. The interviewing program for CATI was the same as that for CAPI. An important new tool for CATI was a call scheduling system. This system took care of proper delivering busy numbers (try again shortly), no answers (try again later), appointments, etc.

By the very early 1990s, nearly all household surveys of Statistics Netherlands had become CAPI or CATI surveys. Surveys using paper forms had almost become extinct. Table 1.3 lists all major and regular household surveys at that time together with their mode of interviewing.

TABLE 1.3 Household surveys carried out by Statistics Netherlands in the early 1990s

Survey	Mode	Interviews per year
Survey on Quality of Life	CAPI	7,500
Health Survey	CAPI	6,200
Day Recreation Survey	CAPI	36,000
Crime Victimisation Survey	CAPI	8,000
Labour Force Survey	CAPI	150,000
Car Use Panel	CATI	8,500
Consumer Sentiments Survey	CATI	24,000
Social-Economic Panel	CATI	5,500
School Career Survey	CATI	4,500
Mobility Survey	CATI/CADI	20,000
Budget Survey	CADI	2,000

In the middle of the 1990s, the MS-DOS operating system on microcomputers was replaced by Windows. This marked the start of the use of graphical user interfaces. Early versions of the Internet browser Internet Explorer were included in this operating system.

Blaise 4 was the first production version of Blaise for Windows released in 1998. When more and more people and companies were connected to the Internet, web surveys became a popular mode of data collection among researchers. The main reasons of this popularity were the high response speed, the possibility to provide feedback to respondents about the meaning of questions and possible errors, and the freedom for the respondents to choose their own moment to fill in the questionnaire.

The graphical user interface offered many more possibilities for screen layout. Figure 1.10 gives an example of a screen of the Blaise 4 CAPI program.

Since respondents are familiar with browsers from all their other activities on the Internet, there was no need to explain the graphical user interface.

The possibility to conduct web surveys was included in version 4.6 of Blaise released in 2003. The respondent completes the questionnaire online allowing continuous interaction between the computer of the respondent and the software on the Internet server.

The Internet questionnaire is divided into pages. Each page may contain one or more questions. After the respondent has answered all questions on a page, the answers are submitted to the Internet server. The answers are checked; a new page is returned to the respondent. The contents of this page may depend on the answers to previous questions.

Figures 1.11 and 1.12 show an example of the same page of a web survey when using Blaise 5. In this case, the page contains only one question. The first page will be displayed when using a tablet, and the second page will be displayed when using a smartphone.

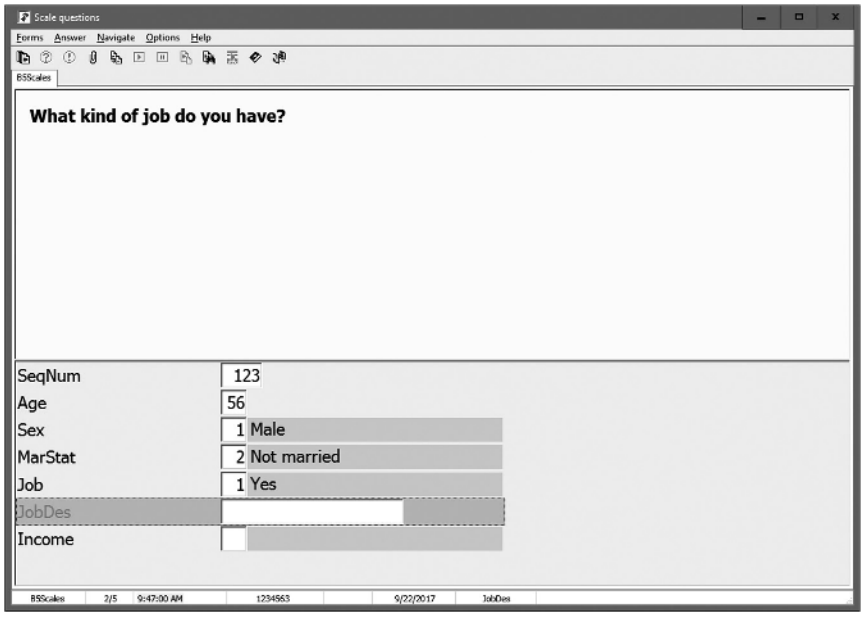

FIGURE 1.10 **The screen of a CAPI program in Blaise 4**

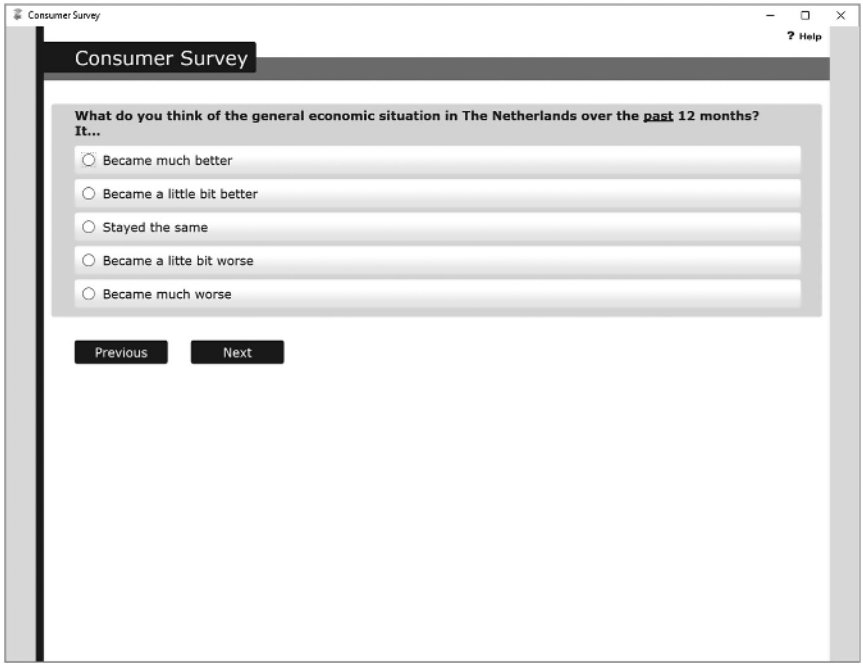

FIGURE 1.11 **The screen of a Blaise 5 web survey on a tablet**

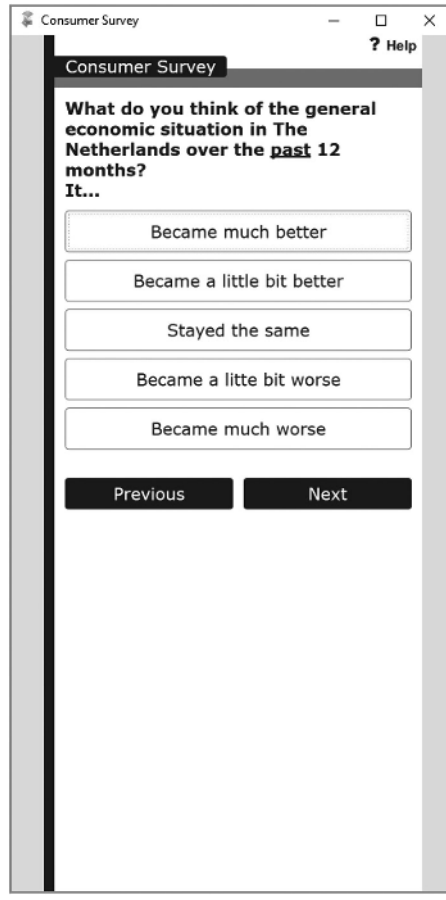

FIGURE 1.12 **The screen of a Blaise 5 web survey on a smartphone**

The Blaise 5 system implements a number of source code features (*Languages, Modes, Roles, and SpecialAnswers*) that specifically address challenges listed above. It also implements a cross-platform layout designer, templates, and cross-platform settings that handle presentation and operability issues. Finally, Blaise 5 allows the institute to combine these features in as many ways as suits its survey program and population.

1.4 Summary

Web surveys are a next step in the evolution process of survey data collection. Collecting data for compiling statistical overviews is already very old, almost as old as mankind. All through history, statistics have been used by rulers of countries to

take informed decisions. However, new developments in society always have had their impact on the way the data were collected for these statistics.

For a long period, until the year 1895, statistical data collection was based on complete enumeration of populations. The censuses were mostly conducted to establish the size of the population, to determine tax obligations of the people, and to measure the military strength of the country.

The first ideas about sampling emerged around 1895. There was a lot of discussion between 1895 and 1934 about how samples should be selected: by means of probability sampling or some other sample selection technique. By 1934 it was clear that only surveys based on probability sampling could provide reliable and accurate estimates. Such surveys were accepted as a scientific method of data collection.

Somewhere in the 1970s another significant development started. The fast development of microcomputers made it possible to introduce CAI. This made survey data collection faster, cheaper, and easier; it also increased data quality. It was time in which acronyms like CATI and CAPI emerged.

The next major development was the creation of the Internet around 1982. When more and more persons and companies got access to the Internet, it became possible to use this network for survey data collection. The first Internet surveys where e-mail surveys. In 1989 the World Wide Web was developed. In the middle of the 1990s, web surveys became popular.

Web surveys are attractive because they allow for simple, fast, and cheap access to large groups of potential respondents. There are, however, also potential methodological problems. There are ample examples of web surveys that are not based on probability sampling. It is not always easy to distinguish good from bad surveys. Attention to the methodological aspects is important both to run web surveys and to use web survey data.

The diffusion of mobile devices, especially smartphones, offers a recent attractive tool to reach interviewee for mobile web surveys, i.e., surveys where the contacted unit can receive and respond using either desk and portable computer or mobile devices. There are however methodological problems to be considered when applying mobile web surveys.

Current digital environment and technology trends are providing a huge amount of data about most phenomena. These data are available on the web and are based on the automatic collection on everything that people do; they are usually called big data; they are not subject to statistical classification criteria and to statistical treatment for representativity. However, they look like an attractive source of data to practitioners and researchers. They are wondering if big data could substitute web surveys to provide information for social and economic decision making. They will not substitute surveys; the message is that the two sources are complementary, but they require the researcher to consider the existing methodological problems. The big data offer a challenging opportunity to revise the role and questions faced from the surveys and to integrate web survey data with other data sources.

KEY TERMS

Blaise: A software package for computer-assisted interviewing and survey processing developed by Statistics Netherlands.

Census: A way of gathering information about a population in which every element in the population has to complete a questionnaire form.

Computer-assisted interviewing (CAI): A form of interviewing in which the questionnaire is not printed on paper. Questions are asked by a computer program.

Computer-assisted personal interviewing (CAPI): A form of face-to-face interviewing in which interviewers use a laptop computer to ask the questions and to record the answers.

Computer-assisted self-administered questionnaires (CASAQ): A form of data collection in which respondents complete the questionnaires on their own computer or device. See also CASI.

Computer-assisted self-interviewing (CASI): A form of data collection in which respondents complete the questionnaires on their own computer or device. See also CASAQ.

Computer-assisted telephone interviewing (CATI): A form of telephone interviewing in which interviewers use a computer to ask the questions and to record the answers.

E-mail survey: A form of data collection via the Internet in which respondents are sent a questionnaire that is part of the body text of an e-mail. The questionnaire is returned by e-mail after answering the questions in the text.

Face-to-face interviewing: A form of interviewing where interviewers visit the homes of the respondents (or another location convenient for the respondent). Together, the interviewer and the respondent complete the questionnaire.

Mail survey: A form of data collection where paper questionnaire forms are sent to the respondents. After completion of the questionnaires, they are returned to the research organization.

Mobile web survey: Self-administered surveys that can be conducted over mobile web-capable devices. They are similar to web surveys, but they have also unique features, such as administration on small screens and keyboards, different navigation, and reaching respondents in various situations, factors that can affect response processes.

Probability sampling: A form of sampling where selection of elements is a random process. Each element must have a positive and known probability of selection.

Purposive sampling: A form of non-probability sampling in which the selection of the sample is based on the judgment of the researcher as to which elements best fit the criteria of the study.

Quota sampling: A form of purposive sampling in which elements are selected from the population in such a way that the distribution of some auxiliary variables matches the population distribution of these variables.

Random digit dialing (RDD): A form of sample selection for a telephone survey where random telephone numbers are generated by some kind of computer algorithm.

Representative method: A methods proposed by Anders Kiaer in 1896 to select a sample from a population in such a way that it forms a "miniature" of the populations.

Straw poll: An informal survey conducted to measure a general feeling of a population. Sample selection is such that it usually does not allow concluding about the population as a whole.

Survey: A way of gathering information about a population in which only a sample of elements from the population has to complete a questionnaire form.

Telephone interviewing: A form of interviewing in which interviewers call selected persons by telephone. If contact is made with the proper person and this person wants to cooperate, the interview is started and conducted over the telephone.

Web scraping: Is data scraping; it is used for extracting data from websites. Note that data scraping is a technique in which a computer program extracts data from human-readable output coming from another program.

Web survey: A form of data collection via the Internet in which respondents complete the questionnaires on the World Wide Web. The questionnaire is accessed by means of a link to a web page.

EXERCISES

Exercise 1.1 Which of the following options is not an advantage of computer-assisted interviewing (CAI) as compared with traditional modes of data collection?

a. Data quality is higher due to included checks.

b. The software is in charge of routing through the questionnaire.

c. CAI leads to higher response rates.

d. Data are processed quicker.

Exercise 1.2 What is an advantage of an e-mail survey over a traditional mail survey?

a. Data quality is higher due to included checks.

b. There is less under-coverage.

c. Response rates are higher.

d. It has better facilities for navigation through the questionnaire.

Exercise 1.3 Why were the first surveys on the Internet e-mail surveys and not web surveys?

a. E-mail surveys were cheaper.

b. The World Wide Web did not exist yet.

c. E-mail surveys are more user friendly.

d. E-mail surveys require less data communication over the Internet.

Exercise 1.4 When should the form-based approach be preferred over the question-by-question approach in a web survey?

a. The questionnaire is very long.

b. The questionnaire contains route instructions and edits.

c. All questions fit on one screen.

d. The survey is a business survey.

Exercise 1.5 Which of the four features is typically and advantage of web surveys?

a. There is no under-coverage.

b. The sample size is always large.

c. A survey can be designed and conducted very quickly.

d. Accurate estimates can always be computed.

Exercise 1.6 How to avoid the problem of under-coverage in a general population web survey?

a. Conduct a mixed-mode survey.

b. Increase the sample size.

c. Conduct a self-selection web survey.

d. Replace the web survey by an e-mail survey.

Exercise 1.7 Why source integration is an interesting perspective?

a. To optimize information.

b. To conduct mixed-mode surveys.

c. To totally avoid web surveys.

d. To run only paper surveys.

REFERENCES

Barcaroli, G., Bianchi, G., Bruni, R., Nurra, A., Salamone, S., & Scarnò, M. (2016), Machine learning and statistical inference: the case of Istat survey on ICT. *Proceeding of the Italian Statistical Society Conference, SIS,* Salerno.

Barcaroli, G., Nurra, A., Salamone, S., Scannapieco, M., Scarnò, M., & Summa, D. (2015), Internet as a Data Source in the Istat Survey on ICT in Enterprises. *Austrian Journal of Statistics*, 44, pp. 31–43. doi:org/10.17713/ajs.v44i2.53.

Beręsewicz, M., Lehtonen, R., Reis, F., Di Consiglio, L., & Karlberg, M. (2018), *An Overview of Methods for Treating Selectivity in Big Data Sources*. Publication Office of the European Union, Luxembourg.

Bethlehem, J. G. (1987), The Data Editing Research Project of the Netherlands Central Bureau of Statistics. *Proceedings of the Third Annual Research Conference of the US Bureau of the Census*, U.S. Bureau of the Census, Washington, DC, pp. 194–203.

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

Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook on Nonresponse in Household Surveys*. Wiley, Hoboken, NJ.

Bethlehem, J. G. & Hofman, L. P. M. (2006), Blaise—Alive and Kicking for 20 Years. *Proceedings of the 10th International Blaise Users Conference*, Arnhem, the Netherlands, pp. 61–86.

Bethlehem, J. G. & Stoop, I. A. L. (2007), Online Panels—A Theft of Paradigm? The Challenges of a Changing World. *Proceedings of the Fifth International Conference of the Association of Survey Computing*, Southampton, U.K., pp. 113–132.

Beukenhorst, D. & Wetzels, W. (2009), *A Comparison of Two Mixed-mode Designs of the Dutch Safety Monitor: Mode Effects, Costs, Logistics*. Technical paper DMH 206546, Statistics Netherlands, Methodology Department, Heerlen, the Netherlands.

Biffignandi, S. & Signorelli, S. (2016), From Big Data to Information: Statistical Issues Through Examples, in Studies. In: Gaul, W., Vichi, M., & Weihs, C. (eds.), *Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin.

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.

Bycroft, C. (2010), *Integrated Household Surveys: A Approach*. Statistics New Zealand, Wellington, New Zealand.

CBS, (1948), *Enige Beschouwingen over Steekproeven*. Reprint from: *Statistische en Economische Onderzoekingen* 3, Statistics Netherlands, The Hague, the Netherlands.

Clayton, R. L. & Werking, G. S. (1998), Business Surveys of the Future: The World Wide Web as a Data Collection Methodology. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls, W. L., & O'Reilly, J. (eds.), *Computer Assisted Survey Information Collection*. Wiley, New York.

Cobben, F. & Bethlehem, J. G. (2005), *Adjusting Under-coverage and Non-response Bias in Telephone Surveys*. Discussion Paper 05006, Statistics Netherlands, Voorburg/Heerlen, the Netherlands.

Cochran, W. G. (1953), *Sampling Techniques*. Wiley, New York.

Couper, M. P. (2013), Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, 7(3), pp. 145–156.

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.

Couper, M. P., Blair, J., & Triplett, T. (1999), A Comparison of Mail and E-mail for a Survey of Employees in U.S. Statistical Agencies. *Journal of Official Statistics*, 15, pp. 39–56.

Couper, M. P. & Nicholls, W. L. (1998), The History and Development of Computer Assisted Survey Information Collection Methods. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls, W. L. & O'Reilly, J. (eds.), *Computer Assisted Survey Information Collection*. Wiley, New York.

de Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21, pp. 233–255.

Deming, W. E. (1950), *Some Theory of Sampling*. Wiley, New York.

Den Dulk, K. & Van Maarseveen, J. (1990), The Population Censuses in The Netherlands. In: Maarseveen, J. V. & Gircour, M. (eds.), *A Century of Statistics, Counting, Accounting and Recounting in The Netherlands*. Statistics Netherlands, Voorburg, the Netherlands.

Dillman, D. A. (2007), *Mail and Internet Surveys: The Tailored Design Method*. Wiley, Hoboken, NJ.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 4th Edition. Wiley, New York.

Graunt, J. (1662), *Natural and Political Observations upon the Bills of Mortality*. Martyn, London, U. K.

Hansen, M. H., Hurvitz, W. N., & Madow, W. G. (1953), *Survey Sampling Methods and Theory*. Wiley, New York.

Horvitz, D. G. & Thompson, D. J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, pp. 663–685.

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, Oslo, Norway.

Kiesler, S. & Sproul, L. S. (1986), Response Effects in the Electronic Survey. *Public Opinion Quarterly*, 50, pp. 402–413.

Laplace, P. S. (1812), *Théorie Analytique des Probabilités. Oevres Complètes*, Vol. 7. Gauthier-Villar, Paris, France.

Lienhard, J. H. (2003), *The Engines of Our Ingenuity, An Engineer Looks at Technology and Culture*. Oxford University Press, Oxford, U.K.

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.

NIPO (1946), Eerste Telefonische Enquête in Nederland verricht door NIPO. *De Publieke Opinie*, 1e jaargang, No. 4.

O'Connell, P. L. (1998), Personal Polls Help the Nosy Sate Curiosity. *New York Times*, June 18.

Pierzchala, M. (2006), *Disparate Modes and Their Effect on Instrument Design*. Paper presented at the 2006 International Blaise Users Conference, Papendal, the Netherlands (at http://www.blaiseusers.org/2006/Papers/207.pdf).

Pierzchala, M. (2016), *Blaise 5—Is Worth the Wait*. Paper presented at the 2016 International Blaise Users Conference, The Hague, the Netherlands (at http://blaiseusers.org/2016/papers/plen2_3.pdf).

Quetelet, L. A. J. (2010), *Lettre à S.A.R. le Duc Régant de Saxe Coburg et Gotha sur la Théorie des Probabilités, Appliquée aux Sciences Morales at Politiques (1846)*. Kessinger Pub Co., Brussels, Belgium.

Quetelet, L. A. J. (2012), *Sur l'Homme et le Développement de ses Facultés, Essai de Physique Sociale (Edit. 1835)*. Hachette Livre-BNF, Paris, France.

Righi, P., Barcaroli, G., & Golini, N. (2017), Quality Issues When Using Big Data in Official Statistics. In: Petrucci, A. & Verde, R. (eds.) *IProceedings of the Conference Statistics and Data Science: New Challenges, New Generations*, FUP Scientific Cloud for Books.

Roos, M., Jaspers, L., & Snijkers, G. (1999), *De Conjunctuurtest via Internet*. Report H4350-99-GWM, Statistics Netherlands, Data Collection Methodology Department, Heerlen, the Netherlands.

Roos, M. & Wings, H. (2000), Blaise Internet Services Put to the Test: Web-surveying the Construction Industry. *Proceedings of the 6th International Blaise Users Conference*, Kinsale, Ireland.

Saris, W. E. (1998), Ten Years of Interviewing Without Interviewers: The Telepanel. In: Couper, M. P., Baker, R. P., Bethlehem, J. G., Clark, C. Z. F., Martin, J., Nicholls II, W. L., & O'Reilly, J. M. (eds.), *Computer Assisted Survey Information Collection*. Wiley, New York, pp. 409–430.

Schaefer, D. R. & Dillman, D. A. (1998), Development of a Standard E-mail Methodology: Results of an Experiment. *Public Opinion Quarterly*, 62, pp. 378–397.

Snijkers, G., Tonglet, J., & Onat, E. (2004), *Projectplan Pilot e-PS*. Internal Report H3424-04-BOO, Development and Support Department, Division of Business Statistics, Statistics Netherlands, Heerlen, the Netherlands.

Snijkers, G., Tonglet, J., & Onat, E. (2005), *Naar een Elektronische Vragenlijst voor Productiestatistieken*. Internal Report, Development and Support Department, Division of Business Statistics, Statistics Netherlands, Heerlen, the Netherlands.

Utts, J. M. (1999), *Seeing Through Statistics*. Duxburry Press, Belmont, CA.

Wells, C. & Thorson, K. (2015), Combining Big Data and Survey Techniques to Model Effects of Political Content Flows in Facebook. *Social Science Computer Review*, 35, pp.1–20.

Witte, J. C., Amoroso, L. M., & Howard, P. E. N. (2000), Method and Representation in Internet-based Survey Tools. *Social Science Computer Review*, 18, pp. 179–195.

Yates, F. (1949), *Sampling Methods for Censuses and Surveys*. Charles Griffin & Co, London, U.K.