Chapter 1

The Cornerstones of Survey Research

Edith D. de Leeuw
Joop J. Hox
Department of Methodology & Statistics, Utrecht University

Don A. Dillman
Washington State University

1.1 INTRODUCTION

The idea of conducting a survey is deceptively simple. It involves identifying a specific group or category of people and collecting information from some of them in order to gain insight into what the entire group does or thinks; however, undertaking a survey inevitably raises questions that may be difficult to answer. How many people need to be surveyed in order to be able to describe fairly accurately the entire group? How should the people be selected? What questions should be asked and how should they be posed to respondents? In addition, what data collection methods should one consider using, and are some of those methods of collecting data better than others? And, once one has collected the information, how should it be analyzed and reported? Deciding to do a survey means committing oneself to work through a myriad of issues each of which is critical to the ultimate success of the survey.

Yet, each day, throughout the world, thousands of surveys are being undertaken. Some surveys involve years of planning, require arduous efforts to select and interview respondents in their home and take many months to complete and many more months to report results. Other surveys are conducted with seemingly lightning speed as web survey requests are transmitted simultaneously to people regardless of their location, and completed surveys start being returned a few minutes later; data collection is stopped in a few days and results are reported minutes afterwards. Whereas some surveys use only one mode of data collection such as the telephone, others may involve multiple modes, for example, starting with mail, switching to telephone, and finishing up with face-to-face interviews. In addition, some surveys are quite simple and inexpensive to do, such as a mail survey of members of a small professional association. Others are incredibly complex, such as a survey of the general public across all countries of the European Union in which the same questions need to be answered in multiple languages by people of all educational levels.

In the mid-twentieth century there was a remarkable similarity of survey procedures and methods. Most surveys of significance were done by face-to-face interviews in most countries in the world. Self-administered paper surveys, usually done by mail, were the only alternative. Yet, by the 1980s the telephone had replaced face-to-face interviews as the dominate survey mode in the United States, and in the next decade telephone surveys became the major data collection method in many countries. Yet other methods were emerging and in the 1990s two additional modes of surveying—the Internet and responding by telephone to prerecorded interview questions, known as Interactive Voice Response or IVR, emerged in some countries. Nevertheless, in some countries the face-to-face interview remained the reliable and predominantly used survey mode.

Never in the history of surveying have their been so many alternatives for collecting survey data, nor has there been so much heterogeneity in the use of survey methods across countries. Heterogeneity also exists within countries as surveyors attempt to match survey modes to the difficulties associated with finding and obtaining response to particular survey populations.

Yet, all surveys face a common challenge, which is how to produce precise estimates by surveying only a relatively small proportion of the larger population, within the limits of the social, economic and technological environments associated with countries and survey populations in countries. This chapter is about solving these common problems that we described as the cornerstones of surveying. When understood and responded to, the cornerstone challenges will assure precision in the pursuit of one's survey objectives.

1.2 WHAT IS A SURVEY?

A quick review of the literature will reveal many different definitions of what constitutes a survey. Some handbooks on survey methodology immediately describe the major components of surveys and of survey error instead of giving a definition (e.g., Fowler, Gallagher, Stringfellow, Zalavsky Thompson & Cleary, 2002, p. 4; Groves, 1989, p. 1), others provide definitions, ranging from concise definitions (e.g., Czaja & Blair, 2005, p. 3; Groves, Fowler, Couper, Lepkowski, Singer & Tourangeau, 2004, p. 2; Statistics Canada, 2003, p. 1) to elaborate descriptions of criteria (Biemer & Lyberg, 2003, Table 1.1). What have these definitions in common? The survey research methods section of the American Statistical Association provides on its website an introduction (Scheuren, 2004) that explains survey methodology for survey users, covering the major steps in the survey process and explaining the methodological issues. According to Scheuren (2004, p. 9) the word survey is used most often to describe a method of gathering information from a sample of individuals. Besides sample and gathering information, other recurring terms in definitions and descriptions are systematic or organized and quantitative. So, a survey can be seen as a research strategy in which quantitative information is systematically collected from a relatively large sample taken from a population.

Most books stress that survey methodology is a science and that there are scientific criteria for survey quality. As a result, criteria for survey quality

have been widely discussed. One very general definition of quality is fitness for use. This definition was coined by Juran and Gryna in their 1980s book on quality planning and analysis, and has been widely quoted since. How this general definition is further specified depends on the product that is being evaluated and the user. For example, quality can be focusing on construction, on making sturdy and safe furniture, and on testing it. Like Ikea, the Swedish furniture chain, that advertised in its catalogs with production quality and gave examples on how a couch was tested on sturdiness. In survey statistics the main focus has been on accuracy, on reducing the mean squared error or MSE. This is based on the Hansen and Hurwitz model (Hansen, Hurwitz, & Madow, 1953; Hansen, Hurwitz, & Bershad, 1961) that differentiates between random error and systematic bias, and offers a concept of total error (see also Kish, 1965), which is still the basis of current survey error models. The statistical quality indicator is thus the MSE: the sum of all squared variable errors and all squared systematic errors. A more modern approach is total quality, which combines both ideas as Biemer and Lyberg (2003) do in their handbook on survey quality. They apply the concept of fitness for use to the survey process, which leads to the following quality requirements for survey data: accuracy as defined by the mean squared error, timeliness as defined by availability at the time it is needed, and accessibility, that is the data should be accessible to those for whom the survey was conducted.

There are many stages in designing a survey and each influences survey quality. Deming (1944) already gave an early warning of the complexity of the task facing the survey designer, when he listed no less than thirteen factors that affect the ultimate usefulness of a survey. Among those are the relatively well understood effects of sampling variability, but also more difficult to measure effects. Deming incorporates effects of the interviewer, method of data collection, nonresponse, questionnaire imperfections, processing errors and errors of interpretation. Other authors (e.g., Kish, 1965, see also Groves, 1989) basically classify threats to survey quality in two main categories, for instance differentiating between errors of nonobservation (e.g., nonresponse) and observation (e.g., in data collection and processing). Biemer and Lyberg (2003) group errors in sampling error and nonsampling error. Sampling error is due to selecting a sample instead of studying the whole population. Nonsampling errors are due to mistakes and/or system deficiencies, and include all errors that can be made during data collection and data processing, such as coverage, nonresponse, measurement, and coding error (see also Lyberg & Biemer, Chapter 22).

In the ensuing chapters of this handbook we provide concrete tools to incorporate quality when designing a survey. The purpose of this chapter is to sensitize the reader to the importance of designing for quality and to introduce the methodological and statistical principles that play a key role in designing sound quality surveys.

A useful metaphor is the design and construction of a house. When building a house, one carefully prepares the ground and places the cornerstones. This is the foundation on which the whole structure must rest. If this foundation is not designed with care, the house will collapse or sink in the unsafe, swampy underground as many Dutch builders have experienced in the past. In the same way, when designing and constructing a survey, one should also lay a well thought-out foundation. In surveys, one starts with preparing the underground

by specifying the concepts to be measured. Then these clearly specified concepts have to be translated, or in technical terms, operationalized into measurable variables. Survey methodologists describe this process in terms of avoiding or reducing specification errors. Social scientists use the term construct validity: the extend to which a measurement method accurately represents the intended construct. This first step is conceptual rather than statistical; the concepts of concern must be defined and specified. On this foundation we place the four cornerstones of survey research: coverage, sampling, response, and measurement (Salant & Dillman, 1994; see also Groves, 1989).

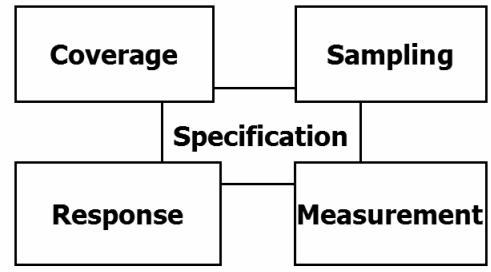

Figure 1.1. The cornerstones of survey research

Figure 1.1 provides a graphical picture of the cornerstone metaphor. Only when these cornerstones are solid, high quality data are collected, which can be used in further processing and analysis. In this chapter we introduce the reader to key issues in survey research.

1.3. BREAKING THE GROUND: SPECIFICATION OF THE RESEARCH AND THE SURVEY QUESTIONS

The first step in the survey process is to determine the research objectives. The researchers have to agree on a well-defined set of research objectives. These are then translated into a set of key research questions. For each research question one or more survey questions are then formulated, depending on the goal of the study. For example, in a general study of the population one or two general questions about well-being are enough to give a global indication of well-being. On the other hand, in a specific study of the influence of social networks on feelings of well-being among the elderly a far more detailed picture of well-being is needed and a series of questions has to be asked, each question measuring a specific aspect of well-being. These different approaches are illustrated in the text boxes noted later.

Example General Well-being Question (Hox, 1986)

Taking all things together, how satisfied or dissatisfied are you with life in general?

- ❑ VERY DISSATISFIED
- ❑ DISSATISFIED
- ❑ NEITHER DISSATISFIED, NOR SATISFIED
- ❑ SATISFIED
- ❑ VERY SATISFIED

Examples General + Specific Well-being Questions (Hox, 1986)

Taking all things together, how satisfied or dissatisfied are you with *life in general*?

- ❑ VERY DISSATISFIED
- ❑ DISSATISFIED
- ❑ NEITHER DISSATISFIED, NOR SATISFIED
- ❑ SATISFIED
- ❑ VERY SATISFIED

Taking all things together, how satisfied or dissatisfied are you with *the home in which you live*?

- ❑ VERY DISSATISFIED
- ❑ DISSATISFIED
- ❑ NEITHER DISSATISFIED, NOR SATISFIED
- ❑ SATISFIED
- ❑ VERY SATISFIED

Taking all things together, how satisfied or dissatisfied are you with *your health*?

Taking all things together, how satisfied or dissatisfied are you with *your social contacts*?

Survey methodologists have given much attention to the problems of formulating the actual questions that go into the survey questionnaire (cf. Fowler & Cosenza, Chapter 8). Problems of question wording, questionnaire flow, question context, and choice of response categories have been the focus of much attention. Much less attention has been directed at clarifying the problems that occur *before* the first survey question is committed to paper: the process that leads from the theoretical construct to the prototype survey item (cf. Hox, 1997). Schwarz (1997) notes that large-scale survey programs often involve a large and heterogeneous group of researchers, where the set of questions finally agreed upon is the result of complex negotiations. As a result, the concepts finally adopted for research are often vaguely defined.

When thinking about the process that leads from theoretical constructs to survey questions, it is useful to distinguish between conceptualization and operationalization. Before questions can be formulated, researchers must decide which concepts they wish to measure. They must define they intend to measure by naming the concept, describing its properties and its scope, and defining important subdomains of its meaning. The subsequent process of operationalization involves choosing empirical indicators for each concept or each subdomain. Theoretical concepts are often referred to as 'constructs' to emphasize that they are theoretical

concepts that have been invented or adopted for a specific scientific purpose (Kerlinger, 1986). Fowler and Cosenza's (Chapter 8) discussion of the distinction between constructs and survey questions follows these line of reasoning.

To bridge the gap between theory and measurement, two distinct research strategies are advocated: a theory driven or top down strategy, which starts with constructs and works toward observable variables and a data driven or bottom up strategy, which starts with observations and works towards theoretical constructs (cf. Hox & De Jong-Gierveld, 1990). For examples of such strategies we refer to Hox (1997).

When a final survey question as posed to a respondent fails to ask about what is essential for the research question, we have a specification error. In other words, the construct implied in the survey question differs from the intended construct that should be measured. This is also referred to as a measurement that has low construct validity. As a result, the wrong parameter is estimated and the research objective is not met. A clear example of a specification error is given by Biemer and Lyberg (2003, p. 39). The intended concept to be measured was "…the value of a parcel of land if it were sold on a fair market today." A potential operationalization in a survey question would be "For what price would you sell this parcel of land?" Closer inspection of this question reveals that this question asks what the parcel of land is subjectively worth to the farmer. Perhaps it is worth so much to the farmer that she/he would never sell it at all.

There are several ways in which one can investigate whether specification errors occur. First of all, the questionnaire outline and the concept questionnaire should always be thoroughly discussed by the researchers, and with the client or information users, and explicit checks should be made whether the questions in the questionnaire reflect the study objectives. In the next step, the concept questionnaire should be pretested with a small group of real respondents, using so called cognitive lab methods. These are qualitative techniques to investigate whether and when errors occur in the question-answer process. The first step in the question answer process is understanding the question. Therefore, the first thing that is investigated in a pretest is if the respondents understand the question and the words used in the question as intended by the researcher. Usually questions are adapted and/or reformulated, based on the results of questionnaire pretests. For a good description of pretesting, methods, see Campanelli Chapter 10. Whenever a question is reformulated, there is the danger of changing its original (intended) meaning, and thus introducing a new specification error. Therefore, both the results of the pretests and the final adapted questionnaire should again be thoroughly discussed with the client.

1.4. PLACING THE CORNERSTONES: COVERAGE, SAMPLING, NONRESPONSE, AND MEASUREMENT

As noted earlier, specification of the research question and the drafting of prototype survey questions are conceptual rather than statistical; it concerns the

construct validity of the measurement. In other words, does the question measure what it is supposed to measure, does it measure the intended theoretical construct (Cronbach & Meehl, 1955). In contrast, the sources of data collection error summarized in our four cornerstones can be assessed statistically by examining the effect they have on the precision of the estimates. Three of the four cornerstones refer explicitly to the fact that surveys typically collect data from a sample, a fraction of the population of interest. *Coverage error* occurs when some members of the population have a zero probability of being selected in the survey sample. For example, the sample list (frame) may fail to cover all elements of the population to which one wants to generalize results. *Sampling error* occurs because only a subset of all elements (people) in the population is actually surveyed. Sampling error is statistically well understood provided that probability samples are used: in general the amount of sampling error is a direct function of the number of units included the final sample. For a clear discussion of coverage and sampling, see Lohr (Chapter 6). *Nonresponse error* occurs when some of the sampled units do not respond and when these units differ from those who do and in a way relevant to the study. For an introduction into nonresponse and nonresponse error, see Lynn (Chapter 3). The last cornerstone is *measurement error*, which occurs when a respondent's answer to a question is inaccurate, departs from the "true" value (see also Hox, Chapter 20).

A perfect survey would minimize all four sources of errors. Coverage error is avoided when every member of the population has a known and nonzero chance of being selected into the survey. Sampling error is reduced simply by sampling enough randomly selected units to achieve the precision that is needed. Nonresponse error is avoided if everyone responds or if the respondents are just like the nonrespondents in terms of the things we are trying to measure. Measurement error can be prevented by asking clear questions; questions that respondents are capable and willing to answer correctly. In the survey design stage the methodological goal is to prevent or at least reduce potential errors; in the analysis stage the statistical goal is to adjust the analysis for errors in such a way that correct (i.e., unbiased and precise) results are produced. The methodological survey literature suggests a variety of methods for reducing the sources of survey error; however, one should keep in mind that there is more than one source of error and that one has to compromise and choose when attempting to reduce total survey error. And, do this all within a workable budget too; or as Lyberg and Biemer put it in Chapter 22: "the challenge in survey design is to achieve an optimal balance between survey errors and costs." In the remainder we discuss the four cornerstones in more detail and relate these to specific chapters in this book.

1.4.1. Coverage and Coverage Error

When doing a survey one has an intended population in mind: the target population. To draw a sample from the target population, a sample frame is needed. This can be a list of target population members, for instance, a list of all members of a certain organization, or the register of all inhabitants of a certain

city. But it may also be a virtual list, or an algorithm, such as in area probability sampling or in Random Digit Dialing (RDD) sampling (cf. Lohr, Chapter 6 on coverage and sampling, and Steeh, Chapter 12 on RDD). In area probability sampling, the population is divided into clusters based on geographical proximity, and then specific areas are selected. In RDD, random telephone numbers are generated using an algorithm that conforms to properties of valid telephone numbers in the country that is being investigated. Frame coverage errors occur when there is a mismatch between the sampling frame and the target population. In other words when there is no one-to-one correspondence between the units in the frame and the units in the target population.

The most common form of coverage error is undercoverage, that is, not all units of the target population are included in the sampling frame. A clear example of undercoverage is persons with an unlisted phone number when the sampling frame is the telephone book. Another form of coverage error is overcoverage; here a unit from the target population appears more than once in the sampling frame. Duplications like this can occur when a sampling frame results from the combination of several lists. For example, on one list a woman is listed under her maiden name, and on a second list under her married name. If these lists are combined, the same person is listed under two different entries. Another example is surveys that use mobile (cell) telephones; these overcover persons who own more than one phone. A third type of coverage error is caused by erroneous inclusions in the frame. For example, a business number is included on a list with household phone numbers.

As a final example, consider the case of web surveys. A common way to attract respondents to a web survey is placing a link to the survey on a popular web site. Basically, this means that the researcher has no control over who responds to the questionnaire. Coverage error for web surveys is related to two different causes (cf. Ramos, Sevedi, & Sweet, 1998). First, it is the respondent who has to make contact with the data collection program. In a web survey, this requires access to a computer and the Internet, plus some degree of computer skill. Individuals who lack these are not covered. In addition, interviewing software is in general not hardware or software independent. Screens look differently in different resolutions, or when different browsers are used to access the survey website, and some combinations of hardware and software may make the survey website inaccessible to some users, resulting in coverage error. For an overview of different types of web surveys and their potential for errors, see Lozar Manfreda and Vehovar (Chapter 14).

The availability of comprehensive lists or algorithms that cover the population differs widely depending on the target population, but also on the country. For instance, in countries like Denmark and The Netherlands the national statistical agency has access to the population registry (see also Bethlehem Chapter 26). This makes it possible for the national statistical agency to draw a probability sample not only of the general population, but also to draw specific subsamples. Some countries have good lists of mobile phone users, whereas others do not. In some areas, the telephone system has a well-defined structure of used and unused number banks, which makes it possible to generate random telephone numbers with good coverage properties. In most areas, the telephone system does not have such a structure or several competing

telephone systems are in use, which makes generating random telephone numbers more difficult (cf. Steeh, Chapter 12).

Web surveys are a special challenge to survey methodologists, because the coverage problem is large and difficult to solve. There are no lists of the population that can be used to draw samples with known properties. Email addresses have no common structure that can be used to generate random addresses similar to the way random telephone numbers are generated in RDD. Finally, the often-used volunteer samples are convenience samples, for which coverage cannot be determined (cf. Lozar Manfreda & Vehovar, Chapter 14).

1.4.2. Sampling and Sampling Error

Sampling error occurs because only a sample of the population is investigated instead of the whole population. Sampling and sampling error is treated by Lohr (Chapter 6). Based on the values for the variables in the *probability* sample, the value for the population is estimated using statistical theory. When simple random sampling is used, standard statistical techniques can be used; however, when more complicated sampling schemes are used, such as cluster sampling or stratification, the standard statistical techniques do not provide accurate *p*-values and confidence intervals and more complicated statistical techniques should be used. Methods for analyzing complex survey designs are discussed by Stapleton in Chapter 18.

Sampling error can be controlled by drawing samples that are large enough to produce the precision wanted. Table 1.1 gives an indication of the number of respondents needed for estimated percentages with a specified precision (e.g., Devore & Peck, 2005, pp. 377–378).

Table 1.1 Precision: Number of respondents needed for percentage estimates within 95 percent Confidence Interval (C.I.).

Number of respondents	Width of 95% C.I.
96	± 10%
384	± 5%
1537	± 2.5%
9604	± 1%

Base percentage 50%, 95% Confidence Interval based on normal approximation

The main point of Table 1.1 is that a large precision requires very large samples. The rule of thumb is that to decrease the sampling errors by half we need a completed sample that is four times as large.

The most important issue about sampling is that if our sample is *not* a probability sample, statistical inference is not appropriate. The difference between probability and nonprobability sampling is that nonprobability sampling does *not* use a *random* selection procedure. This does not necessarily mean that nonprobability samples are unrepresentative of the population; however, it does mean that nonprobability samples cannot depend upon statistical probability theory. With a probabilistic sample, we know the probability that we represent the population well and therefore we can estimate confidence intervals and significance tests. With a nonprobability sample, we

may or may not represent the population well, but it is not appropriate to apply statistical inference to generalize to a general population. At best, we can use statistical inference to assess the precision with which we can generalize to a population consisting of whoever responded. Whether this is representative for any general population is beyond statistical inference.

1.4.3 Response and Nonresponse Error

Nonresponse is the inability to obtain data for all sampled units on all questions. There are two types of nonresponse in surveys: *unit nonresponse* and *item nonresponse*. Unit nonresponse is the failure to obtain any information from an eligible sample unit. Unit nonresponse can be the result of noncontact or refusal. Lynn (Chapter 3) provides an extensive overview on nonresponse and nonresponse error; for a discussion of nonresponse error in cross-cultural studies, see Couper and de Leeuw (2003); for statistical adjustment and weighting see Biemer and Christ (Chapter 16). Item-nonresponse or item missing data refers to the failure to obtain information for one or more questions in a survey, given that the other questions are completed. For an introduction see de Leeuw, Hox, and Huisman (2003), for statistical approaches to deal with missing data see Chapter 18 by Rässler, Rubin, and Schenker.

Nonresponse error is a function of the response rate and the differences between respondents and nonrespondents. If nonresponse is the result of a pure chance process, in other words if nonresponse is completely at random, then there is no real problem. Of course, the realized sample is smaller, resulting in larger confidence intervals around estimators. But the conclusions will not be biased due to nonresponse. Only when respondents and nonrespondents do differ from each other on the variables of interest in the study, will there be a serious nonresponse problem. The nonresponse is then *selective* nonresponse and certain groups may be underrepresented. In the worst case, there is a substantial association between the nonresponse and an important variable of the study causing biased results. A classic example comes from mobility studies: people who travel a lot are more difficult to contact for an interview on mobility than people who travel rarely. Thus, selective nonresponse caused by specific noncontacts leads to an underestimate of mobility. For more examples, see Lynn (Chapter 3).

Two main approaches are used to cope with nonresponse: *reducing* and *adjusting*. Nonresponse reduction applies strategies that, in general, reduce the number of noncontacts and refusals. Causes of noncontact depend on the specific survey design. For instance, in face-to-face surveys, noncontact can be the result of the inability of the interviewer to reach the respondent within the allotted number of contact attempts. Increasing the number of contact attempts not only increases the number of contacted and thus the response rate, but also the costs. Varying the days and times at which contact is attempted also increases the response rate, without affecting the cost as much. In mail and Internet surveys, noncontacts can be the result of undeliverable mailings due to errors in the address list. Tools to reduce refusals also depend on the data collection mode used. For instance, interview surveys may use specially trained interviewers to convert refusals, while mail and Internet surveys have to rely on

incentives or special contacts to counteract explicit refusals. For more detail, see Lynn (Chapter 3).

Nonresponse adjustment refers to statistical adjustments that are applied after the data are collected. If the difference between the respondents and the nonrespondents is known, for instance because we can compare certain characteristics of the respondents to known population values, statistical weighting can be used to make the sample resemble the population with respect to these characteristics. The problem with statistical adjustment is that usually only simple respondent attributes such as age, sex, and education can be used to weigh the sample. This improves the representativeness of the sample with respect to the variables of central substantive interest only if these variables are related to the attributes used in the weighting scheme. Biemer and Christ discuss weighting for survey data in detail in Chapter 17.

Finally, nonresponse figures should be clearly reported in surveys. This often takes the form of a response rate figure. When reporting response rates it is important to state how the response rate was calculated. For details of response rate calculation and a description of sources of nonresponse, see the brochure on standard definitions of the American Association for Public Opinion Research (AAPOR). A regularly updated version and an online response rate calculator can be found on the AAPOR website (www.aapor.org).

1.4.4 Measurement and Measurement Error

Measurement error is also called error of observation. Measurement errors are associated with the data collection process itself. There are three main sources of measurement error: the questionnaire, the respondent, and the method of data collection. When interviewers are used for data collection, the interviewer is a fourth source of error.

A well-designed and well-tested questionnaire is the basis for reducing measurement error. The questions in the questionnaire must be clear, and all respondents must be able to understand the terms used in the same way. With closed questions, the response categories should be well defined, and exhaustive. When a question is not clear, or when the response categories are not clearly defined, respondents will make errors while answering the question or they do not know what to answer. When the data are collected through interviews, interviewers will then try to help out, but in doing this they can make errors too and introduce additional interviewer error (Fowler, 1995). Therefore, improving the *questionnaire* is a good start to improve the total survey quality. For a good introduction into designing and writing effective questions, see Fowler and Cosenza (Chapter 8). It should be emphasized that even carefully designed questionnaires may contain errors and that a questionnaire should always be evaluated and pretested before it may be used in a survey. In Chapter 10 Campanelli provides the reader with information about the different methods for testing survey questions and gives practical guidelines on the implementation of each of the methods.

Respondents can be a source of error in their own right when they provide incorrect information. This may be unintentional, for instance when a respondent does not understand the question or when a respondent has difficulty

remembering an event. But a respondent can also give incorrect information on purpose, for instance when sensitive questions are asked (see also Lensvelt-Mulders, Chapter 23). Measurement errors that originate from the respondent are beyond the control of the researcher. A researcher can only try to minimize respondent errors by making the respondent's task as easy and as pleasant as possible. In other words, by writing clear questions that respondents are willing to answer. In Chapter 2, Schwarz, Knäuper, Oyserman, and Stich describe how respondents come up with an answer and review the cognitive and communicative processes underlying survey responses.

The *method of data collection* can be a third source of measurement error. In Chapter 7 of this book, de Leeuw describes the advantages and disadvantages of major data collection techniques. One of the key differences between survey modes is the way in which certain questions can be asked. For instance, in a telephone interview respondents have to rely on auditive cues only: they only hear the question and the response categories. This may cause problems when a long list of potential answers has to be presented. Dillman, in Chapter 9 on the logic and psychology of questionnaire design, describes mode differences in questionnaire design and proposes a unified or uni mode design to overcome differences between modes. This is of major importance when mixed-mode designs are used, either within one survey, or in longitudinal studies (e.g., panel surveys see also Chapter 25 by Sikkel & Hoogendoorn), or between surveys as can be the case in cross-national and comparative studies in which one mode (e.g., telephone) is used in one country an another mode (e.g., face-to-face interviews) is used in another. For important issues in comparative survey research, see Harkness (Chapter 4); for more detail on the challenges of mixed mode surveys, see De Leeuw, Dillman, and Hox (Chapter 16).

A second major difference between modes is the presence versus the absence of an interviewer. There may be very good reasons to choose a method without interviewers and leave the locus of control with the respondents, such as ensuring more privacy and more time to reflect for respondents. Self-administered questionnaires in general are described by De Leeuw and Hox in Chapter 13; technological innovations are described by Lozar Manfreda and Vehovar in Chapter 14 on Internet Surveys and by Miller Steiger and Conroy in Chapter 15 on Interactive Voice Response. On the other hand, using interviewers also has many positive points, especially when very complex questionnaires are used or when special tasks have to be performed. As Loosveldt states in Chapter 11: "…the task of the interviewer is more comprehensive and complex than merely asking questions and recording the respondent's answer. Interviewers implement the contact procedure, persuade the respondents to participate, clarify the respondent's role during the interview and collect information about the respondent."

However, when an interviewer is present, the interviewer can be a source of error too. Interviewers may misinterpret a question, may make errors in administering a questionnaire, or in registering the answers. When posing the question, interviewers may unintentionally change its meaning. By giving additional information or explaining a misunderstood word, they may inappropriately influence a respondent. Even the way interviewers look and dress may influence a respondent in a face-to-face interview. Selecting and

training interviewers carefully helps reducing interviewer related errors. For more details, see Chapter 23 on interviewer training by Lessler, Eyerman, and Wang. Interviewers can make genuine mistakes, but they also may intentionally cheat. Interviewers have been known to falsify data, or skip questions to shorten tedious interviews. Monitoring interviewers helps to reduce this. Having a quality controller listening in on telephone interviewers is a widely used method. In face-to-face interviews, recordings can be made and selected tapes can be checked afterwards. Special verification contacts or re-interviews may be used to evaluate interviewer performance in large-scale face-to-face surveys (cf. Lyberg & Biemer, Chapter 22; Japec, 2005, p. 24).

1.5 FROM DATA COLLECTION TO ANALYSIS: HOW THE FOUNDATION AFFECTS THE STRUCTURE

There are several ways in which the design of a survey and the precise data collection procedure affects the subsequent data analysis stage. These also involve the four cornerstones. The most direct influence is the actual *sampling* procedure that is used. As mentioned earlier, standard statistical procedures assume that the data are a simple random sample from the population. In most surveys, other sampling schemes are used because these are more efficient or less expensive, for instance cluster sampling or stratification. When these sampling schemes are used, the analysis must employ special statistical methods (see also Stapleton, Chapter 17). Similarly, when weighting (cf. Biemer & Christ, Chapter 16) is used to compensate for different inclusion probabilities, either by design or because of nonresponse problems, special statistical methods must be used. Standard statistical packages may or may not include these methods. For instance, the package SPSS (version 15 and higher) can analyze complex survey data with weights and complicated sampling schemes, but it includes only selected statistical analyses for such data. The other procedures in SPSS can include weighting, but do not correct the standard errors for the effects of weighting, which produces incorrect statistical tests.

A less obvious way in which the survey design affects the data analysis lies in the adjustment for the combination of coverage error and nonresponse. These may result in data that are not representative for the population, and the most often-used adjustment method is weighting on respondent characteristics for which the population values are known. For more detail, see Biemer and Christ (Chapter 16). Two issues are important here. First, statistical adjustment aims at producing unbiased estimates of population parameters when selection probabilities are not equal; however, no amount of statistical cleverness restores information that we have failed to collect. So, prevention by reducing the problem in the data collection phase is important. Second, the quality of the adjustment depends strongly on the amount and quality of background information that we have available to construct the weights. Collecting this information requires careful planning in the design phase. Auxiliary variables must be included for which the population values are known, for instance for a sample from the general population via the national statistical agency, or for samples from a special population via an existing registry. Because the use of

registries is regulated by privacy concerns, in the latter case it may be necessary to obtain prior permission. For more on privacy and ethics in survey research, see Singer (Chapter 5). Finally, to be able to use the information, it is crucial that the data collection procedure uses the same wording and response categories that were used to collect the known population data (cf. Dillman, Chapter 9). Preferably, the same method of data collection should be used, to prevent confounding of selection and measurement errors.

A special case of nonresponse is the failure to obtain information on some of the questions, which leads to incomplete data for some of the respondents. Just as is the case with unit-nonresponse discussed earlier, prevention and the collection of auxiliary information is important with item missing data too (see also de Leeuw, Hox, & Huisman, 2003). The next step is statistical adjustment. In Chapter 19, Rässler, Rubin, and Schenker discuss concepts regarding mechanisms that create missing data, as well as four commonly used approaches to deal with (item) missing data.

Measurement errors, that is discrepancies between the measurement and the true value, influence the analysis in more subtle ways. Again, prevention is the best medicine. Measurement errors originate from the question wording and the questionnaire, from the survey method and the interviewer, from the respondents and from complex interactions between these. Many decisions in the survey design phase have the potential to affect measurement error (cf. Biemer & Lyberg, Chapter 22). Prevention rest on the application of known best practices in survey design; this assumes that these are well documented (cf. Mohler, Pennel, & Frost, Chapter 21). Another important step in reducing measurement error as far as possible is thorough pretesting of the survey instrument before it is actually used (cf. Campanelli, Chapter 10). In the analysis phase, some adjustments for the effect of measurement errors can be made; Hox discusses this in Chapter 19. Adjustments for measurement errors can be made when multi-item scales are used, or if auxiliary information is available about the amount of measurement error in specific variables. Again, to be able to adjust in the analysis phase, the design of the survey must make sure that the necessary information is available.

1.6 CAN WE AFFORD IT: BALANCING DESIGN FEATURES AND SURVEY QUALITY

Earlier we discussed the foundation of survey research: breaking the ground (specification) and placing the four cornerstones (coverage, sampling, nonresponse, and measurement). The same fundamental quality criteria are discussed in quality handbooks. For instance, in Eurostat's 2000 publication on the assessment of quality in statistics, the first quality criterion is the relevance of the statistical concept. A statistical product is relevant if it meets user's needs. This implies that user's needs must be established at the start. The concept of relevance is closely related to the specification problem and the construct validity of measurement. Did we correctly translate the substantive research question into a survey question? If not, we have made a specification error, and the statistical product does not meet the needs of the users. Almost all handbooks on survey

statistics mention *accuracy* of the estimate as quality criterion. Accuracy depends on all four cornerstones and is discussed at length earlier in this chapter. But, there are additional criteria for quality as well. Biemer and Lyberg (2003) stress the importance of timeliness defined as available at the time it is needed, and accessibility, that is the data should be accessible to those for whom the survey was conducted. Eurostat (2000) distinguishes seven distinct dimensions of statistical quality, adding a.o. comparability, meaning that it should be possible to make reliable comparisons across time and across space. Comparability is extremely important in cross-cultural and cross-national studies (see also Harkness, Chapter 4). For a discussion of quality and procedures for quality assurance and quality control, see Lyberg and Biemer (Chapter 22).

Both Biemer and Lyberg's (2003) quality concepts and Eurostat's (2000) dimensions go beyond the foundation and cornerstones described earlier in this chapter, and are relevant for the quality of the entire survey process and the data it produces. Their criteria were developed mainly for use in large scale survey organizations and governmental statistical offices, but survey quality and quality assurance is an issue that also applies to smaller scale surveys, where the survey researcher is also the survey user. It does not matter if it is a small scale survey or a large survey, whether the survey is using paper and pencil or high technology, quality can and should be built into all surveys. For procedures for quality assessment, see Lyberg and Biemer (Chapter 22).

To come back to the metaphor of building a house: there are many different ways to build a good, quality house. But, there is also a large variety in types of houses, ranging from a simple summer cottage to a luxurious villa, from a houseboat to a monumental 17th century house at a canal, from a working farm to a dream palace. What is a good house depends on the needs of the resident, what is a good survey depends on the survey user (cf. Dippo, 1997). The research objectives determine the population under study and the types of questions that should be asked. Privacy regulations and ethics may restrict the design; other practical restriction may be caused by available time and funds. Countries and survey organizations may differ in available resources, such as skilled labor, administrative capacities, experience with certain procedures or methods, computer hardware and software. It is clear that survey methodologists must balance survey costs and available resources against survey errors, and that any actual survey will be the result of methodological compromises. Surveys are a complex enterprise and many aspects must be considered when the goal is to maximize data quality with the available resources and within a reasonable budget of time and costs.

Finally, surveys are carried out in a specific cultural context, which may also affect the way these aspects influence the survey quality. Survey methodologists need to take this into account when designing a survey. For instance, when a telephone (or Internet) survey is contemplated for an international study, it is important to understand how telephones and Internet are viewed in the different cultures included in the survey. Is it a personal device, such as mobile telephones? Is it a household device, as landline telephones mostly are? Or is it a community device, with one (mobile) telephone or Internet connection shared by an entire village? Survey design means that costs and quality must be optimized, and in a global world this

means that they must be optimized within the bounds of cultural and technological resources and differences.

1.7 CONTENTS OF THIS BOOK

The goal of this book is to introduce the readers to the central issues that are important for survey quality, to discuss the decisions that must be made in designing and carrying out a survey, and to present the current methodological and statistical knowledge about the consequences of these decisions for the survey data quality.

The first section of the book, *Foundations*, is a broad introduction in survey methodology. In addition to this introduction, it contains chapters on the psychology of asking questions, the problem of nonresponse, issues and challenges in international surveys, and ethical issues in surveys.

The second section, *Design*, presents a number of issues that are vital in designing a quality survey. It includes chapters on coverage and sampling, choosing the method of data collection, writing effective questions, constructing the questionnaire, and testing survey questions.

The third major section, *Implementation*, discusses the details of a number of procedures to carry out a survey. There are chapters on face-to-face interviews, telephone interviews, self-administered questionnaires, Internet surveys and Interactive Voice Response surveys. Finally, there is a chapter on the challenges that result when different data collection modes are mixed within a survey.

The fourth section, *Data analysis*, discusses a number of statistical subjects that are especially important in analyzing survey data. These include chapters on constructing adjustment weights, analyzing data from complex surveys, coping with incomplete data (item nonresponse), and accommodating measurement errors. The final section, *Special issues*, contains a number of special interest topics for quality surveys. It includes chapters on survey documentation, quality assurance and quality control, interviewer training, collecting data on sensitive topics, and panel surveys including access panels. The final chapter introduces collecting survey-type data without asking questions of respondents, by combining and integrating existing information.

GLOSSARY OF KEY CONCEPTS

Construct validity. The extend to which a measurement instrument measures the intended construct and produces an observation distinct from that produced by a measure of a different construct.

Coverage error. Coverage errors occur when the operational definition of the population includes an omission, duplication, or wrongful inclusion of an element in the population. Omissions lead to undercoverage, and duplications and wrongful inclusions lead to overcoverage.

Measurement error. The extent to which there are discrepancies between a measurement and the true value, that the measurement instrument is designed to

measure. Measurement error refers to both variance and bias, where variance is random variation of a measurement and bias is systematic error. There are a number of potential sources; for example, measurement error can arise from the respondent, questionnaire, mode of data collection, interviewer, and interactions between these.

Nonresponse error. Nonresponse is the failure to collect information from sampled respondents. There are two types of nonresponse: unit nonresponse and item nonresponse. Unit nonresponse occurs when the survey fails to obtain any data from a unit in the selected sample. Item nonresponse (incomplete data) occurs when the unit participates but data on particular items are missing. Nonresponse leads to nonresponse error if the respondents differ from the nonrespondents on the variables of interest.

Sampling error. Error in estimation due to taking a sample instead of measuring every unit in the sampling frame. If probability sampling is used then the amount of sampling error can be estimated from the sample.

Specification error. Specification error occurs when the concept measured by a survey question and the concept that should be measured with that question differ. When this occurs, there is low construct validity.

Chapter 2

The Psychology of Asking Questions

Norbert Schwarz
University of Michigan

Bärbel Knäuper
McGill University

Daphna Oyserman
University of Michigan

Christine Stich
McGill University

2.1 INTRODUCTION

Over the last two decades, psychologists and survey methodologists have made considerable progress in understanding the cognitive and communicative processes underlying survey responses, increasingly turning the "art of asking questions" (Payne, 1951) into an applied science that is grounded in basic psychological research. This chapter reviews key lessons learned from this work (for more extended reviews see Schwarz 1999a; Sirken, Hermann, Schechter, Schwarz, Tanur, & Tourangeau, 1999; Sudman, Bradburn, & Schwarz 1996; Tourangeau, Rips, & Rasinski 2000). We focus on how features of the research instrument shape respondents' answers and illustrate how the underlying processes can change as a function of respondents' age and culture. We first address respondents' tasks and subsequently discuss how respondents make sense of the questions asked. Next, we review how respondents answer behavioral questions and relate these questions to issues of autobiographical memory and estimation. Finally, we address attitude questions and review the conditions that give rise to context effects in attitude measurement.

2.2 RESPONDENTS' TASKS

It is now widely recognized that answering a survey question involves several tasks. Respondents first need to understand the question to determine which information they are asked to provide. Next, they need to recall relevant information from memory. When the question is an opinion question, they will

rarely find a ready-for-use answer stored in memory. Instead, they need to form a judgment on the spot, based on whatever relevant information comes to mind at that time. When the question pertains to a behavior, respondents need to retrieve relevant episodes. Unless the behavior is rare and important, this is a difficult task and respondents typically have to rely on inference and estimation strategies to arrive at an answer. Once respondents have formed a judgment in their own minds, they can rarely report it in their own words. Instead, they need to format it to fit the response alternatives provided by the researcher. Finally, respondents may hesitate to communicate their private judgment, because of social desirability and self-presentation. If so, they may edit their judgment before conveying it to the researcher. Accordingly, understanding the question, recalling information, forming a judgment, formatting the judgment to fit the response alternatives, and editing the final answer are the major steps of the question answering process (see Strack & Martin, 1987; Tourangeau, 1984).

Unfortunately, respondents' performance at each of these steps is highly context dependent. From a psychological perspective, this context dependency is part and parcel of human cognition and communication, in daily life as in survey interviews. From a survey methods perspective, however, it presents a formidable problem: To the extent that the answers provided by the sample are shaped by the research instrument, they do not reflect the opinions or behaviors of the population to which the researcher wants to generalize. Complicating things further, a growing body of findings suggests that the underlying processes are age- and culture-sensitive, resulting in differential context effects that can thwart straightforward comparisons across cohorts and cultures.

2.3 UNDERSTANDING THE QUESTION

Survey textbooks typically advise researchers to avoid unfamiliar terms and complex syntax (for helpful guidelines see Bradburn, Sudman, & Wansink, 2004). This is good advice, but it misses a crucial point: Language comprehension is not about words per se, but about speaker meaning (Clark & Schober, 1992). Respondents certainly understand the words when asked, "What have you done today?" But to provide a meaningful answer they need to determine which behaviors the researcher might be interested in. For example, should they report that they took a shower, or not? To infer the intended meaning of the question, respondents rely on the tacit assumptions that govern the conduct of conversation in daily life. These assumptions were described by Paul Grice (1975), a philosopher of language, in the form of four maxims: A *maxim of relation* asks speakers to make their contribution relevant to the aims of the ongoing conversation. A *maxim of quantity* requests speakers to make their contribution as informative as is required, but not more informative than is required. A *maxim of manner* holds that a speaker's contribution should be clear rather than obscure, ambiguous or wordy, and a *maxim of quality* requires speakers not to say anything that's false. In short, speakers should try to be informative, truthful, relevant, and clear and listeners interpret the speakers' utterances "on the assumption that they are trying to live up to these ideals" (Clark & Clark, 1977, p. 122).

Respondents bring these tacit assumptions to the research situation and assume that the researcher "chose his wording so they can understand what he meant—and can do so quickly" (Clark & Schober, 1992, p. 27). To do so, they draw on the context of the ongoing conversation to determine the question's intended meaning, much as they would be expected to do in daily life. In fact, reliance on contextual information is more pronounced under the standardized conditions of survey interviews, where a well trained interviewer may merely reiterate the identical question, than under the less constrained conditions of daily life, which allow for mutual clarifications of the intended meaning. The contextual information provided by the researcher includes formal features of the questionnaire, in addition to the specific wording of the question and the content of preceding questions, as a few examples may illustrate (see Clark & Schober, 1992; Schwarz, 1996; Strack, 1994, for reviews).

2.3.1 Response Alternatives

Returning to the previously mentioned example, suppose respondents are asked in an *open response format*, "What have you done today?" To give a meaningful answer, they have to determine which activities may be of interest to the researcher. In an attempt to be informative, they are likely to omit activities that the researcher is obviously aware of (e.g., "I gave a survey interview") or may take for granted anyway (e.g., "I had breakfast"), thus observing the maxim of quantity. But most respondents would endorse these activities if they were included in a list presented as part of a *closed response format.* On the other hand, a closed response format would reduce the likelihood that respondents report any activities omitted from the list (see Schuman & Presser, 1981; Schwarz & Hippler, 1991, for reviews). This reflects that response alternatives convey what the researcher is interested in, thus limiting the range of "informative" answers. In addition, they may remind respondents of material that they may otherwise not consider.

Even something as innocuous as the *numeric values of rating scales* can elicit pronounced shifts in question interpretation. Schwarz, Knäuper, Hippler, Noelle-Neumann, and Clark (1991) asked respondents how successful they have been in life, using an 11-point rating scale with the endpoints labeled "not at all successful" and "extremely successful." To answer this question, respondents need to determine what is meant by "not at all successful"—the absence of noteworthy achievements or the presence of explicit failures? When the numeric values of the rating sale ranged from 0 to 10, respondents inferred that the question refers to different degrees of success, with "not at all successful" marking the absence of noteworthy achievements. But when the numeric values ranged from -5 to +5, with 0 as the middle alternative, they inferred that the researcher had a bipolar dimension in mind, with "not at all successful" marking the opposite of success, namely the presence of failure. Not surprisingly, this shift in the meaning of the verbal endpoint labels resulted in dramatic shifts in the obtained ratings. Whereas 34% of the respondents endorsed a value between 0 and 5 on the 0 to 10 scale, only 13% endorsed one of the formally equivalent values between -5 and 0 on the -5 to +5 scale 0, reflecting that the absence of great success is more common than the presence of failure. Hence, researchers are well advised to match the numeric values to the intended uni- or bipolarity of the scale.

The numeric values of behavioral *frequency scales* can serve a similar function. For example, Schwarz, Strack, Müller, and Chassein (1988) asked respondents to report how often they are angry along a scale that presented either high or low frequency values. As expected, respondents inferred that the question pertains to more intense anger experiences, which are relatively rare, when accompanied by low frequency values, but to mild anger experiences when accompanied by high frequency values. Throughout, respondents assume that the researcher constructs meaningful response alternatives that are relevant to the specific question asked, consistent with Grice's (1975) maxim of relation.

2.3.2 Question Wording

Similar issues apply to question wording. Minor changes in apparently formal features of the question can result in pronounced meaning shifts, as the case of *reference periods* may illustrate. Winkielman, Knäuper, and Schwarz (1998) asked respondents, in an open response format, either how frequently they had been angry last week or last year. Respondents inferred that the researcher is interested in less frequent and more severe episodes of anger when the question pertained to one year rather than to one week—after all, they could hardly be expected to remember minor anger episodes for a one-year period, whereas major anger may be too rare to make a one-week period plausible. Hence, they reported on rare and intense anger for the one year period, but more frequent and less intense anger for the one week period and their examples reflected this differential question interpretation. Accordingly, it is not surprising that reports across different reference periods do not add up—respondents may not even report on the same type of experience to begin with, thwarting comparisons across reference periods.

2.3.3 Question Context

Respondents' interpretation of a question's intended meaning is further affected by the context in which the question is presented. Hence, a question about drugs acquires a different meaning in the context of health versus a crime survey. Not surprisingly, the influence of *adjacent questions* is more pronounced for more ambiguously worded questions, which force respondents to rely on the context information to infer the intended meaning (e.g., Strack, Schwarz, & Wänke, 1991). Survey researchers have long been aware of this possibility (e.g., Payne, 1951). What is often overlooked, however, is that the *researcher's affiliation*, conveyed in the cover letter, may serve a similar function. For example, Norenzayan and Schwarz (1999) observed that respondents provided more personality focused explanations of a behavior when the questionnaire was printed on the letterhead of an "Institute for Personality Research" rather than an "Institute for Social Research." Such differences highlight the extent to which respondents as cooperative communicators attempt to make their answers relevant to the inferred epistemic interest of the researcher (see Schwarz, 1996).

2.3.4 Age-related Differences

Respondents' extensive use of contextual information requires that they hold the question in mind and relate it to other aspects of the questionnaire to determine its intended meaning. This entails considerable demands on respondents' cognitive resources. Given that these resources decline with increasing age (for a review see Park, 1999), we may expect that older respondents are less likely to use, or less successful in using, contextual information at the question comprehension stage. A limited body of findings supports this conjecture. For example, Schwarz, Park, Knäuper, Davidson, and Smith (1998) observed that older respondents (aged over 70) were less likely than younger respondents to draw on the numeric values of rating scales to interpret the meaning of endpoint labels. Similarly, Knäuper (1999a) observed in secondary analyses that question order effects decrease with age, as addressed in the section on attitude questions. Moreover, children and adolescents, whose cognitive capabilities are not yet fully developed, appear to show a similar deficit in incorporating relevant contextual information into survey responding (Borgers, de Leeuw, & Hox, 2000; Fuchs, 2005).

On theoretical grounds, age-related differences in the use of contextual information should be particularly likely in face-to-face and telephone interviews, where respondents can not look back to earlier questions. In contrast, they may be less pronounced in self-administered questionnaires, where respondents can deliberately return to previous questions when they encounter an ambiguous one (Schwarz & Hippler, 1995). If so, age-related differences in the response process may interact with the mode of data collection, further complicating comparisons across age groups.

2.3.5 Implications for Questionnaire Construction

As the preceding examples illustrate, question comprehension is not solely an issue of understanding the literal meaning of an utterance. Instead, it involves extensive inferences about the speaker's intentions to determine the pragmatic meaning of the question. To safeguard against unintended question interpretations and related complications, psychologists and survey methodologists have developed a number of procedures that can be employed in questionnaire pretesting (see Campanelli, chapter 10; Schwarz & Sudman, 1996). These procedures include the extensive use of probes and think-aloud protocols (summarily referred to as cognitive interviewing; e.g., DeMaio & Rothgeb, 1996), detailed coding of interview transcripts (e.g., Fowler & Cannell, 1996), and the use of expert systems that alert researchers to likely problems (e.g., Lessler & Forsyth, 1996). Without such development efforts, respondents' understanding of the questions asked may differ in important ways from what the researcher had in mind.

2.4 REPORTING ON ONE'S BEHAVIORS

Many survey questions pertain to respondents' behaviors, often asking them to report how frequently they engaged in a given behavior during a specified reference period. Ideally, respondents are supposed to determine the boundaries of the reference period and to recall all instances of the behavior within these boundaries to arrive at the relevant frequency. Unfortunately, respondents are usually unable to follow this recall-and-count strategy, unless the behavior is rare and important and the reference period short and recent (Menon, 1994). Instead, respondents will typically need to rely on estimation strategies to arrive at a plausible approximation. Next, we review key aspects of autobiographical memory and subsequently address respondents' estimation strategies.

2.4.1 Autobiographical Memory

Not surprisingly, people forget events in their lives as time goes by, even when the event is relatively important and distinct. For example, Cannell, Fisher, and Bakker (1965) observed that only 3% of their respondents failed to report an episode of hospitalization when interviewed within ten weeks of the event, yet a full 42% did so when interviewed one year after the event. Moreover, when the question pertains to a frequent behavior, respondents are unlikely to have detailed representations of numerous individual episodes of a behavior stored in memory. Instead, the various instances of closely related behaviors blend into one global, knowledge-like representation that lacks specific time or location markers (Linton, 1982; Strube, 1987). As a result, individual episodes of frequent behaviors become indistinguishable and irretrievable. Throughout, the available research suggests that the recall of individual behavioral episodes is largely limited to rare and unique behaviors of considerable importance, and poor even under these conditions.

Complicating things further, our autobiographical knowledge is not organized by categories of behavior (like drinking alcohol) that map easily onto survey questions. The structure of autobiographical memory can be thought of as a hierarchical network that includes extended periods (like "the years I lived in New York") at the highest level of the hierarchy. Nested within this high-order period are lower-level extended events pertaining to this time, like "my first job" or "the time I was married to Lucy." Further down the hierarchy are summarized events, which correspond to the knowledge-like representations of repeated behaviors noted earlier (e.g., "During that time, Lucy and I quarreled a lot"). Specific events, like a particular episode of disagreement, are represented at the lowest level of the hierarchy. To be represented at this level of specificity, however, the event has to be rather unique. As these examples illustrate, autobiographical memory is primarily organized by time ("the years in New York") and relatively global themes ("first job"; "first marriage") in a hierarchical network (see Belli, 1998, for a review). The search for any specific event in this network takes considerable time and the outcome is somewhat haphazard, depending on the entry point into the network at which the search started. Hence, using multiple entry points and forming connections across different periods and themes improves recall.

2.4.2 Facilitating Recall

Drawing on basic research into the structure of autobiographical memory, researchers have developed a number of strategies to facilitate autobiographical recall (for reviews see Schwarz & Oyserman, 2001; Sudman et al., 1996; Schwarz & Sudman, 1994; Tourangeau et al., 2000).

To some extent, researchers can improve the likelihood of accurate recall by restricting the recall task to a short and recent reference period. This strategy, however, may result in many zero answers from respondents who rarely engage in the behavior, thus limiting later analyses to respondents with high behavioral frequencies. As a second strategy, researchers can provide appropriate recall cues. In general, the date of an event is the poorest cue, whereas cues pertaining to what happened, where it happened, and who was involved are more effective (e.g., Wagenaar, 1986). Note, however, that recall cues share many of the characteristics of closed response formats and can constrain the inferred question meaning. It is therefore important to ensure that the recall cues are relatively exhaustive and compatible with the intended interpretation of the question.

Closely related to the provision of recall cues is the *decomposition* of a complex task into several more specific ones. Although this strategy results in reliable increases in reported frequency (e.g., Blair & Burton, 1987; Sudman & Schwarz, 1989), "more" is not always "better" and decomposition does not necessarily increase the accuracy of the obtained reports (e.g., Belli, Schwarz, Singer, & Talarico, 2000). As many studies documented, frequency estimates are regressive and people commonly overestimate low frequencies, but underestimate high frequencies (see Belli et al., 2000 for a review).

In addition, autobiographical recall will improve when respondents are given sufficient time to search memory. Recalling specific events may take up to several seconds and repeated attempts to recall may result in the retrieval of additional material, even after a considerable number of previous trials (e.g., Williams & Hollan, 1981). Unfortunately, respondents are unlikely to have sufficient time to engage in repeated retrieval attempts in most research situations. Moreover, they may often not be motivated to do so even if they had the time. Accordingly, explicitly instructing respondents that the next question is really important, and that they should do their best and take all the time they may need, has been found to improve recall (e.g., Cannell, Miller, & Oksenberg, 1981). Note, however, that it needs to be employed sparingly and may lose its credibility when used for too many questions within an interview.

Although the previously mentioned strategies improve recall to some extent, they fail to take full advantage of what has been learned about the hierarchical structure of autobiographical memory. A promising alternative approach is offered by the *event history calendar* (see Belli, 1998, for a review), which takes advantage of the hierarchically nested structure of autobiographical memory to facilitate recall. To help respondents recall their alcohol consumption during the last week, for example, they may be given a calendar grid that provides a column for each day of the week, cross-cut by rows that pertain to relevant contexts. They may be asked to enter for each day what they did, who they were

with, if they ate out, and so on. Reconstructing the last week in this way provides a rich set of contextual cues for recalling episodes of alcohol consumption.

2.4.3 Estimation Strategies

Given the reviewed memory difficulties, it is not surprising that respondents usually resort to a variety of inference strategies to arrive at a plausible estimate (for a review see Sudman et al., 1996, Chapter 9). Even when they can recall relevant episodic information, the recalled material may not cover the entire reference period or they may be aware that their recall is likely to be incomplete. In such cases, they may base their inferences on the recalled fragments, following a *decomposition* strategy (e.g., Blair & Burton, 1987). In other cases, respondents may draw on *subjective theories* that bear on the behavior in question (for a review see Ross, 1989). When asked about past behavior, for example, they may ask themselves if there is reason to assume that their past behavior was different from their present behavior—if not, they may report their present behavior as an approximation. Schwarz and Oyserman (2001) review these and related strategies. Here, we illustrate the role of estimation strategies by returning to respondents' use of information provided by formal characteristics of the questionnaire.

2.4.4 Response Alternatives

In many studies, respondents are asked to report their behavior by checking the appropriate response alternative on a *numeric frequency scale.* Consistent with Grice's (1975) maxim of relation, respondents assume that the researcher constructed a meaningful scale that is relevant to the task at hand. Specifically, they assume that values in the middle range of the scale reflect the average or "usual" behavior, whereas values at the extremes of the scale correspond to the extremes of the distribution. Given these assumptions, respondents can draw on the range of the response alternatives as a plausible frame of reference in estimating their own behavioral frequency. This results in higher frequency estimates when the scale presents high rather than low frequency values.

For example, Schwarz and Scheuring (1992) asked 60 patients of a German mental health clinic to report the frequency of 17 symptoms along one of the following two scales:

Low Frequency Scale	**High Frequency Scale**
( ) never	( ) twice a month or less
( ) about once a year	( ) once a week
( ) about twice a year	( ) twice a week
( ) twice a month	( ) daily
( ) more than twice a month	( ) several times a day

Across 17 symptoms, 62% of the respondents reported average frequencies of more than twice a month when presented with the high frequency scale, whereas only 39% did so when presented with the low frequency scale, resulting in a mean difference of 23 percentage points. This influence of

frequency scales has been observed across a wide range of different behaviors, including health behaviors, television consumption (e.g., Schwarz, Hippler, Deutsch, & Strack, 1985), sexual behaviors (e.g., Tourangeau & Smith, 1996), and consumer behaviors (e.g., Menon, Rhaghubir, & Schwarz, 1995).

On theoretical grounds, we may expect that the impact of numeric frequency values is more pronounced, the more poorly the behavior is represented in memory, thus forcing respondents to rely on an estimation strategy. Empirically, this is the case. The influence of frequency scales is small when the behavior is rare and important, and hence well represented in memory. Moreover, when a respondent engages in the behavior with high regularity (e.g., every Sunday), its frequency can easily be derived from this rate information, largely eliminating the impact of frequency scales (Menon, 1994; Menon et al., 1995).

2.4.5 Age- and Culture-related Differences in Estimation

Given age-related declines in memory, we may expect that the impact of response alternatives is more pronounced for older than for younger respondents. The available data support this prediction with some qualifications. For example, Knäuper, Schwarz, and Park (2004) observed that the frequency range of the response scale affected older respondents more than younger respondents when the question pertained to mundane behaviors, such as buying a birthday present. On the other hand, older respondents were less affected than younger respondents when the question pertained to the frequency of physical symptoms, which older people are more likely to monitor, resulting in better memory representations.

Similarly, Ji, Schwarz, and Nisbett (2000) observed pronounced cultural differences in respondents' need to estimate. In general, collectivist cultures put a higher premium on "fitting in" than individualist cultures (Oyserman, Coon, & Kemmelmeier, 2002). To "fit in," people need to monitor their own publicly observable behavior as well as the behavior of others to note undesirable deviations. Such monitoring is not required for private, unobservable behaviors. We may therefore expect that public behaviors are better represented in memory for people living in collectivistic rather than individualistic cultures, whereas private behaviors may be equally poorly represented in both cultures. To test these conjectures, Ji and colleagues (2000) asked students in China and the United States to report public and private behaviors along high or low frequency scales, or in an open response format. Replicating earlier findings, American students reported higher frequencies when presented with a high rather than low frequency scale, independent of whether the behavior was private or public. Chinese students' reports were similarly influenced by the frequency scale when the behavior was private, confirming that they relied on the same estimation strategy. In contrast, Chinese students' reports were unaffected by the response format when the behavior was public and hence needed to be monitored to ensure social fit.

As these examples illustrate, social groups differ in the extent to which they pay close attention to a given behavior. These differences in behavioral monitoring, in turn, influence to which extent respondents need to rely on estimation strategies in reporting on their behaviors, rendering them differentially susceptible to contextual influences. Importantly, such differences in respondents'

strategies can result in misleading substantive conclusions about behavioral differences across cultures and cohorts.

2.4.6 Subsequent Judgments

In addition to affecting respondents' behavioral reports, frequency scales can also affect respondents' *subsequent judgments*. For example, respondents who check a frequency of twice a month on one of Schwarz and Scheuring's (1992) scales, shown earlier, may infer that their own symptom frequency is above average when presented with the low frequency scale, but below average when presented with the high frequency scale. Empirically, this is the case and the patients in this study reported higher health satisfaction after reporting their symptom frequencies on the high rather than low frequency scale – even though patients given a high frequency scale had reported a higher absolute symptom frequency to begin with. Again, such scale-induced comparison effects have been observed across a wide range of judgments (see Schwarz, 1999b for a review).

2.4.7 Editing the Answer

After respondents arrived at an answer in their own mind, they need to communicate it to the researcher. At this stage, the communicated estimate may deviate from their private estimate due to considerations of social desirability and self-presentation as already mentioned (see DeMaio, 1984, for a review. Not surprisingly, editing on the basis of social desirability is particularly likely in response to threatening questions and is more pronounced in face-to-face interviews than in self-administered questionnaires, which provide a higher degree of confidentiality. All methods designed to reduce socially desirable responding address one of these two factors. Bradburn et al. (2004) review these methods and provide good advice on their use (see also Lensvelt-Mulders, Chapter 24).

2.4.8 Implications for Questionnaire Construction

In sum, respondents will rarely be able to draw on extensive episodic memories when asked to report on the frequency of mundane behaviors. Instead, they need to rely on a variety of estimation strategies to arrive at a reasonable answer. Which strategy they use is often influenced by the research instrument, as the case of frequency scales illustrates. The most basic way to improve behavioral reports is to ensure that respondents have sufficient time to search memory and to encourage respondents to invest the necessary effort (Cannell et al., 1981). Moreover, it is usually advisable to ask frequency questions in an open response format, such as, "How many times a week do you …? ___ times a week." Although the answers will not be accurate, the open response format will at least avoid the systematic biases associated with frequency scales.

Given these memory problems, researchers are often tempted to simplify the task by merely asking respondents if they engage in the behavior "never," "sometimes," or "frequently." Such *vague quantifiers,* however, are come with their own set of problems (see Pepper, 1981, for a review). For example,

"frequently" suffering from headaches reflects higher absolute frequencies than "frequently" suffering from heart attacks, and “sometimes” suffering from headaches denotes a higher frequency for respondents with a medical history of migraine than for respondents without that history. In general, the use of vague quantifiers reflects the objective frequency relative to respondents' subjective standard, rendering vague quantifiers inadequate for the assessment of objective frequencies, despite their popularity.

2.5 REPORTING ON ONE'S ATTITUDES

Public opinion researchers have long been aware that attitude measurement is highly context dependent. In this section, we address the two dominant sources of context effects in attitude measurement, namely the order in which questions and response alternatives are presented to respondents.

2.5.1 Question Order Effects

Dating back to the beginning of survey research, numerous studies demonstrated that preceding questions can influence the answers given to later questions (see Schuman & Presser, 1981; Schwarz & Sudman, 1992; Sudman et al., 1996; Tourangeau et al., 2000, for reviews). Moreover, when a self-administered questionnaire is used, respondents can go back and forth between questions, occasionally resulting in influences of later questions on responses to earlier ones (e.g., Schwarz & Hippler, 1995).

Question order effects arise for a number of different reasons. First, preceding questions can affect respondents’ inferences about the intended meaning of subsequent questions, as discussed in the section on question comprehension (e.g., Strack, Schwarz, & Wänke, 1991). Second, they can influence respondents’ use of rating scales, resulting in less extreme ratings when a given item is preceded by more extreme ones, which serve as scale anchors (e.g., Ostrom & Upshaw, 1968). Third, they can bring general norms to mind that are subsequently applied to other issues (e.g., Schuman & Ludwig, 1983). Finally, preceding questions can influence which information respondents use in forming a mental representation of the attitude object and the standard against which the object is evaluated.

The accumulating evidence suggests that a differential construal of attitude objects and standards is the most common source of question order effects. Hence, we focus on this aspect by following Schwarz and Bless' (1992a) inclusion/exclusion model, which predicts the direction and size of question order effects in attitude measurement, as well as their generalization across related issues.

2.5.2 Mental Construal

Attitude questions assess respondents’ evaluations of an attitude object. From a psychological perspective, evaluations require two mental representations: A representation of the to-be-evaluated target and a representation of a standard,

against which the target is assessed. Both of these representations are formed on the basis of information that is accessible at the time of judgment. This includes information that may always come to mind when the respondent thinks about this topic (chronically accessible information), as well as information that may only come to mind because of contextual influences, for example information that was used to answer earlier questions (temporarily accessible information). Whereas temporarily accessible information is the basis of most context effects in attitude measurement, chronically accessible information lends some context-independent stability to respondents' judgments.

Independent of whether the information is chronically or temporarily accessible, people truncate the information search as soon as enough information has come to mind to form a judgment with sufficient subjective certainty. Hence, their judgment is rarely based on all information that may bear on the topic, but dominated by the information that comes to mind most easily at that point in time. How this information influences the judgment, depends on how it is used.

2.5.3 Assimilation Effects

Information that is *included* in the temporary representation formed of the target results in *assimilation effects*. That is, including information with positive implications results in a more positive judgment, whereas including information with negative implications results in a more negative judgment. For example, Schwarz, Strack, and Mai (1991) asked respondents to report their marital satisfaction and their general life-satisfaction in different question orders. When the general life-satisfaction question was asked first, it correlated with marital satisfaction $r = .32$. Reversing the question order, however, increased this correlation to $r = .67$. This reflects that the marital satisfaction question brought marriage related information to mind that respondents included in the representation formed of their lives in general. Accordingly, happily married respondents reported higher general life-satisfaction in the marriage-life than in the life-marriage order, whereas unhappily married respondents reported lower life-satisfaction under this condition.

As this pattern indicates, the specific effect of thinking about one's marriage depends on whether it is a happy or unhappy one. Accordingly, no overall mean difference was observed for the sample as a whole, despite pronounced differences in correlation. As a general principle, question order effects are not a function of the preceding question per se, but of the information that the question brings to mind. Hence, pronounced question order effects may occur in the absence of overall mean differences, rendering measures of association more sensitive than examinations of means.

Theoretically, the size of assimilation effects increases with the amount and extremity of the temporarily accessible information, and decreases with the amount and extremity of chronically accessible information, that is included in the representation of the target (e.g., Bless, Schwarz, & Wänke, 2003). To continue with the previously mentioned example, some respondents were asked to report on their job satisfaction, leisure satisfaction, and marital satisfaction prior to reporting on their general life-satisfaction, thus bringing a more varied range of information about their lives to mind. As expected, this decreased the correlation

of marital satisfaction and general life-satisfaction from $r = .67$ to $r = .43$. By the same token, we expect that respondents who are experts on a given issue show less pronounced assimilation effects than novices, because experts can draw on a larger set of chronically accessible information, which in turn reduces the impact of adding a given piece of temporarily accessible information. Note, however, that expert status needs to be defined with regard to the specific issue at hand. Global variables, such as years of schooling, are unlikely to moderate the size of assimilation effects, unless they are confounded with the amount of knowledge regarding the issue under consideration. Accordingly, formal education has been found to show inconsistent relationships with the emergence and size of question order effects (Schuman & Presser, 1981).

2.5.4 Contrast Effects

What has long rendered the prediction of question order effects challenging, is that the same piece of information that elicits an assimilation effect may also result in a *contrast effect.* This is the case when the information is excluded from, rather than included in, the cognitive representation formed of the target (Schwarz & Bless, 1992a). As a first possibility, suppose that a given piece of information with positive (negative) implications is excluded from the representation of the target. If so, the representation contains less positive (negative) information, resulting in a less positive (negative) judgment. For example, the Schwarz et al. (1991) life-satisfaction study included a condition in which the marital satisfaction and life-satisfaction questions were introduced with a joint lead-in that read, "We now have two questions about your life. The first pertains to your marriage and the second to your life in general." This lead-in was designed to evoke the conversational maxim of quantity (Grice, 1975), which enjoins speakers to avoid redundancy when answering related questions. Accordingly, respondents who had just reported on their marriage should now disregard this aspect of their lives when answering the general life-satisfaction question. Confirming this prediction, happily married respondents now reported lower general life-satisfaction, whereas unhappily married respondents reported higher life-satisfaction, indicating that they excluded the positive (negative) marital information from the representation formed of their lives in general. These diverging effects reduced the correlation to $r = .18$, from $r = .67$ when the same questions were asked in the same order without a joint lead-in. Finally, a control condition in which the general life-satisfaction question was reworded to, "Aside from your marriage, which you already told us about, how satisfied are you with your life in general?" resulted in a highly similar correlation of $r = .20$. Such *subtraction based* contrast effects are limited to the specific target (here, one's life in general), reflecting that merely subtracting a piece of information (here, one's marriage) does only affect this specific representation. The size of subtraction based contrast effects increases with the amount and extremity of the temporarily accessible information that is excluded from the representation of the target, and decreases with the amount and extremity of the information that remains in the representation of the target.

As a second possibility, respondents may not only exclude accessible information from the representation formed of the target, but may also use this information in constructing a standard of comparison. If the implications of the

temporarily accessible information are more extreme than the implications of the chronically accessible information used in constructing a standard, this process results in a more extreme standard, eliciting contrast effects for that reason. The size of these comparison based contrast effects increases with the extremity and amount of temporarily accessible information used in constructing the standard or scale anchor, and decreases with the amount and extremity of chronically accessible information used in making this construction. In contrast to subtraction based comparison effects, which are limited to a specific target, comparison based contrast effects generalize to all targets to which the standard is applicable.

As an example, consider the impact of political scandals on assessments of the trustworthiness of politicians. Not surprisingly, thinking about a politician who was involved in a scandal, say Richard Nixon, decreases trust in politicians in general. This assimilation effect reflects that the exemplar is included in the representation formed of the target politicians in general. If the trustworthiness question pertains to a specific politician, however, say Bill Clinton, the primed exemplar cannot be included in the representation formed of the target—after all, Bill Clinton is not Richard Nixon. In this case, Richard Nixon may serve as a standard of comparison, relative to which Bill Clinton seems very trustworthy. Experiments with German exemplars confirmed these predictions (Schwarz & Bless, 1992b; Bless, Igou, Schwarz, & Wänke, 2000): Thinking about a politician who was involved in a scandal decreased the trustworthiness of politicians in general, but increased the trustworthiness of all specific exemplars assessed. In general, the same information is likely to result in assimilation effects in the evaluation of superordinate target categories (which allow for the inclusion of all information pertaining to subordinate categories), but in contrast effects in the evaluation of lateral target categories (which are mutually exclusive).

2.5.5 Determinants of Inclusion/Exclusion

Given the crucial role of inclusion/exclusion operations in the construction of mental representations, it is important to understand their determinants. When thinking about a topic, people generally assume that whatever comes to mind bears on what they are thinking about—or why else would it come to mind now? Hence, the default information is to include information that comes to mind in the representation of the target. This renders assimilation effects more likely than contrast effects. In fact, assimilation effects (sometimes referred to as carry-over effects) dominate the survey literature and many models intended to account for question order effects don't even offer a mechanism for the conceptualization of contrast effects (e.g., Zaller, 1992), which severely limits their usefulness as general theoretical frameworks. Whereas inclusion is the more common default, the exclusion of information needs to be triggered by salient features of the question answering process. The most relevant variables can be conceptualized as bearing on three implicit decisions that respondents have to make with regard to the information that comes to mind.

Some information that comes to mind may simply be irrelevant, pertaining to issues that are unrelated to the question asked. Other information may potentially be relevant to the task at hand and respondents have to decide what to do with it. The first decision bears on why this information comes to

mind. Information that seems to come to mind for the wrong reason, for example because respondents are aware of the potential influence of a preceding question, is likely to be excluded. The second decision bears on whether the information that comes to mind bears on the target of judgment or not. The content of the context question (e.g., Schwarz & Bless, 1992a), the superordinate or lateral nature of the target category (e.g., Schwarz & Bless, 1992b), the extremity of the information (e.g., Herr, 1986), or its representativeness for the target category (e.g., Strack, Schwarz, & Gschneidinger, 1985) are relevant at this stage. Finally, conversational norms of nonredundancy may elicit the exclusion of previously provided information, as seen earlier (Schwarz et al., 1991).

Whenever any of these decisions results in the exclusion of information from the representation formed of the target, it will elicit a contrast effect. Whether this contrast effect is limited to the target, or generalizes across related targets, depends on whether the excluded information is merely subtracted from the representation of the target or used in constructing a standard against which the target is evaluated. Whenever the information that comes to mind is included in the representation formed of the target, on the other hand, it results in an assimilation effect. Hence, the inclusion/exclusion model provides a coherent conceptualization of the emergence, direction, size, and generalization of context effects in attitude measurement (see Schwarz & Bless, 1992a; Sudman et al., 1996, Chapter 5, for more detail).

2.5.6 Age- and Culture-related Differences

To guard against question order effects, survey researchers often separate related questions with buffer items. These buffer items presumably render the previously used information less accessible, thus attenuating the influence of earlier questions (for a review see Wänke & Schwarz, 1997). The same logic suggests that preceding questions should be less likely to influence the judgments of older respondents, due to age-related declines in memory. Empirically this is the case, as Knäuper (1999a) observed in secondary analyses of survey data.

Much as age-related differences in memory performance can elicit age-sensitive context effects, culture-related differences in conversational practice can elicit culture-sensitive context effects. For example, Asian cultures value more indirect forms of communication, which require a higher amount of reading between the lines, based on high sensitivity to subtle conversational cues. Accordingly, Asians are more likely to notice the potential redundancy of related questions, as Haberstroh, Oysermen, Schwarz, Kühnen and Ji (2002) observed in a conceptual replication of the previously mentioned marital satisfaction study (Schwarz et al., 1991) with Chinese respondents. Throughout, such age- and culture-sensitive context effects can invite misleading conclusions about age- and culture-related differences in respondents' attitudes.

2.5.7 Response Order Effects

Another major source of context effects in attitude measurement is the order in which response alternatives are presented. Response order effects are most

reliably obtained when a question presents several plausible response options (see Sudman et al., 1996, chapter 6, for a detailed discussion). Suppose, for example, that respondents are asked in a self-administered questionnaire whether divorce should be easier to obtain or more difficult to obtain. When they first think about the easier option, they may quickly come up with a good reason for making divorce easier and may endorse this answer. But had they first thought about the more difficult option, they might as well have come up with a good reason for making divorce more difficult and might have endorsed that answer. In short, the order in which response alternatives are presented can influence the mental representation that respondents form of the issue (see Sudman et al., 1996, for a more detailed discussion).

Which response alternative respondents are more likely to elaborate on first, depends on the presentation order and mode (Krosnick & Alwin, 1987). In a visual format, like a self-administered questionnaire, respondents think about the response alternatives in the order in which they are presented. In this case, a given alternative is more likely to be endorsed when presented first rather than last, resulting in a *primacy effect*. In an auditory format, like a telephone interview, respondents cannot think about the details until the interviewer has read the whole question. In this case, they are likely to begin with the last alternative read to them, which is still in their ear. Under this format, a given alternative is more likely to be endorsed when presented last rather than first, resulting in a recency effect.

2.5.8 Age-related Differences

On theoretical grounds, we may expect that age-related limitations of working memory capacity further enhance respondents' tendency to elaborate mostly on a single response alternative. Empirically this is the case and an extensive meta-analysis documented that response order effects are more pronounced for older and less educated respondents (Knäuper, 1999b). This age-sensitivity of response order effects can again invite misleading conclusions about cohort differences in the reported attitude, suggesting, for example, that older respondents are more liberal than younger respondents under one order condition, but more conservative under the other (Knäuper, 1999a).

The observation that response order effects increase with age, whereas question order effects decrease with age, also highlights that age-sensitive context effects do indeed reflect age-related differences in cognitive capacity, which can plausibly account for both observations. In contrast, attempts to trace these differences to age-related differences in attitude strength (e.g., Sears, 1986) would suggest that question order and response order effects show parallel age patterns, which is not the case.

2.5.9 Implications for Questionnaire Construction

Human judgment is always context dependent, in daily life as in survey interviews. Although attention to the theoretical principles summarized earlier can help researchers to attenuate context effects in attitude measurement, the best safeguard against misleading conclusions is the experimental variation of question and response order within a survey.

2.6 CONCLUDING REMARKS

Survey researchers have long been aware that collecting data by asking questions is an exercise that may yield many surprises. Since the 1980s, psychologists and survey methodologists have made considerable progress in understanding the cognitive and communicative processes underlying question answering, rendering some of these surprises less surprising than they have been in the past. Yet, this does not imply that we can always predict how a given question would behave when colleagues ask us for advice: In many cases, the given question is too mushy an operationalization of theoretical variables to allow for predictions (although we typically feel we know what would happen if the question were tinkered with, in one way or another, to bring it in line with theoretical models). Nevertheless, the accumulating insights (reviewed in Sudman et al., 1996; Tourangeau et al., 2000) alert us to likely problems and help us in identifying questions and question sequences that need systematic experimental testing before they are employed in a large-scale study.

GLOSSARY OF KEY CONCEPTS

Assimilation effect. A catch-all term for any influence that makes the answers to two questions more similar than they otherwise would be; it does not entail specific assumptions about the underlying process.
Backfire effect. See contrast effect.
Carry-over effect. See assimilation effect.
Context effect. A catch-all term for any influence of the context in which a question is asked; it does not entail specific assumptions about the direction of the effect or the underlying process.
Contrast effect. A catch-all term for any influence that makes the answers to two questions more different than they otherwise would be; it does not entail specific assumptions about the underlying process.
Pragmatic meaning. Refers to the intended (rather than literal or semantic) meaning of an utterance and requires inferences about the speaker's knowledge and intentions.
Primacy effect. A given response alternative is more likely to be chosen when presented at the beginning rather than at the end of a list of response alternatives.
Question order effect. The order in which questions are asked influences the obtained answers; different processes can give rise to this influence.
Recency effect. A given response alternative is more likely to be chosen when presented at the end rather than at the beginning of a list of response alternatives.
Response order effect. The order in which response alternatives are presented influences which alternative is endorsed; see primacy effect and recency effect.
Semantic meaning. Refers to the literal meaning of words. Understanding the semantic meaning is insufficient for answering a question, which requires an understanding of the question's pragmatic meaning.

Chapter 3

The Problem of Nonresponse

Peter Lynn
University of Essex

3.1 INTRODUCTION

Many books about survey sampling show how the precision of survey estimates depends on the sample design; however, this assumes that data are obtained for every unit in the selected sample. This is rarely the case; most surveys experience some nonresponse. Consequently, the sample upon which the estimates are based is not the same as the sample that was originally selected. Obviously, it is smaller. But it may also be different in other important ways that affect the estimates.

It may seem rather negative to be discussing nonresponse so early in this book. We haven't yet begun to discuss how to design or implement a survey and yet we are already talking about failure—failure to collect data from all the units in our sample. But this is a fundamental aspect of survey research. If we cannot successfully collect data from a large proportion of the selected units, then it may be a waste of time carrying out a survey at all. And when the data have been collected and we want to make estimates we need to be able to make allowances for the effect of nonresponse. This requires advance planning—even before the sample has been selected. In this chapter, I try to explain how and why nonresponse occurs, why it is important, and what we can do to minimize any undesirable consequences.

3.2 WHY IS NONRESPONSE IMPORTANT?

Even the most well resourced surveys carried out by experienced survey organizations suffer from nonresponse. The level of nonresponse can vary greatly between surveys, depending on the nature of the sample units, the mode of data collection, the fieldwork procedures used and societal and cultural factors. Some of these factors vary between countries and often lead to response rates differing between countries for the same survey. But whatever the circumstances of your survey, you are almost certain to have some nonresponse.

The principles of statistical inference (see Lohr, Chapter 6) allow us to make inferences about a population of interest, provided that the sample has been selected using a known probability mechanism. In other words, we have to know the selection probability of each unit in our sample. But nonresponse

disturbs the selection probabilities. The probability of a particular unit being in our final responding sample, sometimes referred to as the inclusion probability, is the product of the original selection probability and the probability of the unit responding once selected. Assuming that we have used a probability sampling design, the first of these is known. But the second is not known. The result is that our sample may no longer be representative of the population.

Consider a simple example of a survey of literacy in a small town. Suppose we want to estimate the proportion of adults classified as low ability, based upon a test that will be administered as part of the survey interview (ignore for the moment the fact that the test may not provide a perfectly accurate measure of ability—see Hox, Chapter 20). Imagine that the population of 14,000 adults in the town consists of 8,000 who would be classified as high ability if the test were administered and 6,000 who would be classified as low ability (though of course we would not know this). The sample design is to randomly select one in every 20 adults (see Table 3.1), so we would expect to find approximately 400 high ability and 300 low ability persons in our sample. Suppose however that the low ability persons are less likely to respond to the survey, with a response probability of only 0.60, compared with 0.80 for the high ability persons. This means that we can expect to find 180 low ability persons in the responding sample of 500, so we might estimate the proportion of low ability persons in the population to be 36%, whereas in fact it is 43% (6,000 out of 14,000). But if we were carrying out this survey for real, we might not be aware that our estimate is too low. We would only observe the numbers highlighted in bold in Table 3.1. In the absence of other information, we would have no way of knowing that low ability persons had been less likely to respond to the survey and no reason to adjust our estimate of 36%.

Table 3.1: The effect of nonresponse on a survey of literacy

	High ability	Low ability	Total
Population	8,000	6,000	**14,000**
Selection probability	**1/20**	**1/20**	**1/20**
Expected sample size	400	300	**700**
Response probability	0.80	0.60	**0.714**
Responding sample size	**320**	**180**	**500**

Note: Figures in bold would be known; other figures not

This error in our estimate has been caused by nonresponse. Specifically, it has been caused by the fact that the response probability is associated with the target variable (literacy ability). If nonresponse had happened completely at random, then we would still have expected to find 43% of the responding sample to be low ability. But nonresponse rarely happens completely at random. There are reasons why some units do not respond and those reasons are typically associated with at least some of the survey variables. In our example, it may be that some residents of the town were away in a different location, engaged in seasonal employment, during the survey field work period. If such

people were selected into the sample, it would not have been possible to contact them so they would have been nonrespondents. And if people with low literacy ability were more likely than those with high ability to engage in this seasonal employment, this could lead to exactly the sort of effect shown in Table 3.1.

3.3 HOW DOES NONRESPONSE ARISE?

There are several reasons why nonresponse occurs. If we are to be successful in trying to minimize the extent of nonresponse, we need to understand these reasons and to find ways of combating each of them. A summary classification of reasons for nonresponse appears in Table 3.2. These reflect the stages of the survey data collection process. Once a sample unit is selected, it is first necessary for the data collector to identify the location of that unit. This may prove impossible if, for example, the address information on the sampling frame is incomplete (a). If located successfully, the next step is to make contact with the sample unit. Sometimes, as in the example above, this proves impossible (b). Even if contact is made successfully, it may not prove possible to collect the required data. Reasons for this can be broadly classified into three types: the sample unit may be unwilling to co-operate (c), or unable to co-operate (d), or the data collector and sample unit may be unable to communicate adequately (e). Finally, it sometimes happens that data are successfully collected from the sample unit but subsequently lost–for example if questionnaires go missing in the post or computer files become corrupted (f).

Table 3.2: Reasons for nonresponse

a. Failure of the data collector to locate/identify the sample unit
b. Failure to make contact with the sample unit
c. Refusal of the sample unit to participate
d. Inability of the sample unit to participate (e.g. ill health, absence, etc)
e. Inability of the data collector and sample unit to communicate (e.g. language barriers)
f. Accidental loss of the data/ questionnaire

This simple classification provides a framework for considering reasons for nonresponse but it does not describe the many specific reasons that could apply on any particular survey. Often, reasons for nonresponse will be specific to the topic of the survey, to the types of units from which data are to be collected, and to the way that the survey is designed and carried out. In particular, there are important differences between surveys carried out by face-to-face interviewing, by telephone interviewing, and by self-completion methods. There are also differences between surveys of individuals and households on the one hand and businesses and other establishments on the other. In the case of individuals and households, there is also an important distinction between surveys where the data are collected in the sample member's own home and

surveys where the sample member is responding in a different context or in a particular capacity (e.g., as a user of a particular service or as a visitor to a particular place). Let us consider some common types of survey.

3.3.1 Face-to-face Interview Surveys of Households or Individuals

Many surveys of the household population in a country, region or town are carried out using face-to-face interviews in the respondents' own home. For example, most national statistical offices carry out Labor Force Surveys and Household Budget Surveys in this way. The World Bank's series of Living Standards Measurement Surveys (http://www.worldbank.org/lsms/) are also carried out in this way. The sample is usually selected from a list of either persons or addresses (e.g., a population register, a list of postal addresses, or a list of addresses drawn up in the field as part of the survey preparation phase) and the interviewers' first task is to locate each selected address. They must then make contact with the residents, confirm whether any resident is eligible for the survey, possibly make a random selection of one person to interview, contact the selected person, persuade the person to be interviewed, agree a convenient time and place for the interview, administer the interview, and transmit the data to the survey office. At each stage, nonresponse could occur for each of several reasons. To illustrate this, consider the example of surveys of individuals in the United Kingdom, where a sample of addresses is selected from the Post Office list, and one person is subsequently selected for interview at each address. Surveys that use this design include the British Crime Survey, the British Social Attitudes Survey and the UK part of the European Social Survey. Similar designs can be found in several other countries. The fieldwork process is summarized in Figure 3.1. The shaded boxes indicate nonresponse outcomes.

The first stage of the process is to mail an advance letter (or prenotification letter) to each selected address. This notifies the residents that an interviewer will be visiting soon, provides some basic information about the survey, and provides contact details for the survey organization in case the recipient has queries or concerns. Having received this letter, some sample members contact the survey organization to indicate that they do not wish to participate in the survey. Where possible, the survey organization attempts to persuade these sample members to allow the interviewer to visit and to explain the survey in more detail, emphasizing that they will still have the opportunity to decline to take part at that stage if they wish. But this is not always successful; some sample members insist that they do not want an interviewer to visit. These cases are typically referred to as office refusals, as they are refusals noted in the survey office, before the interviewer has had a chance to influence the outcome.

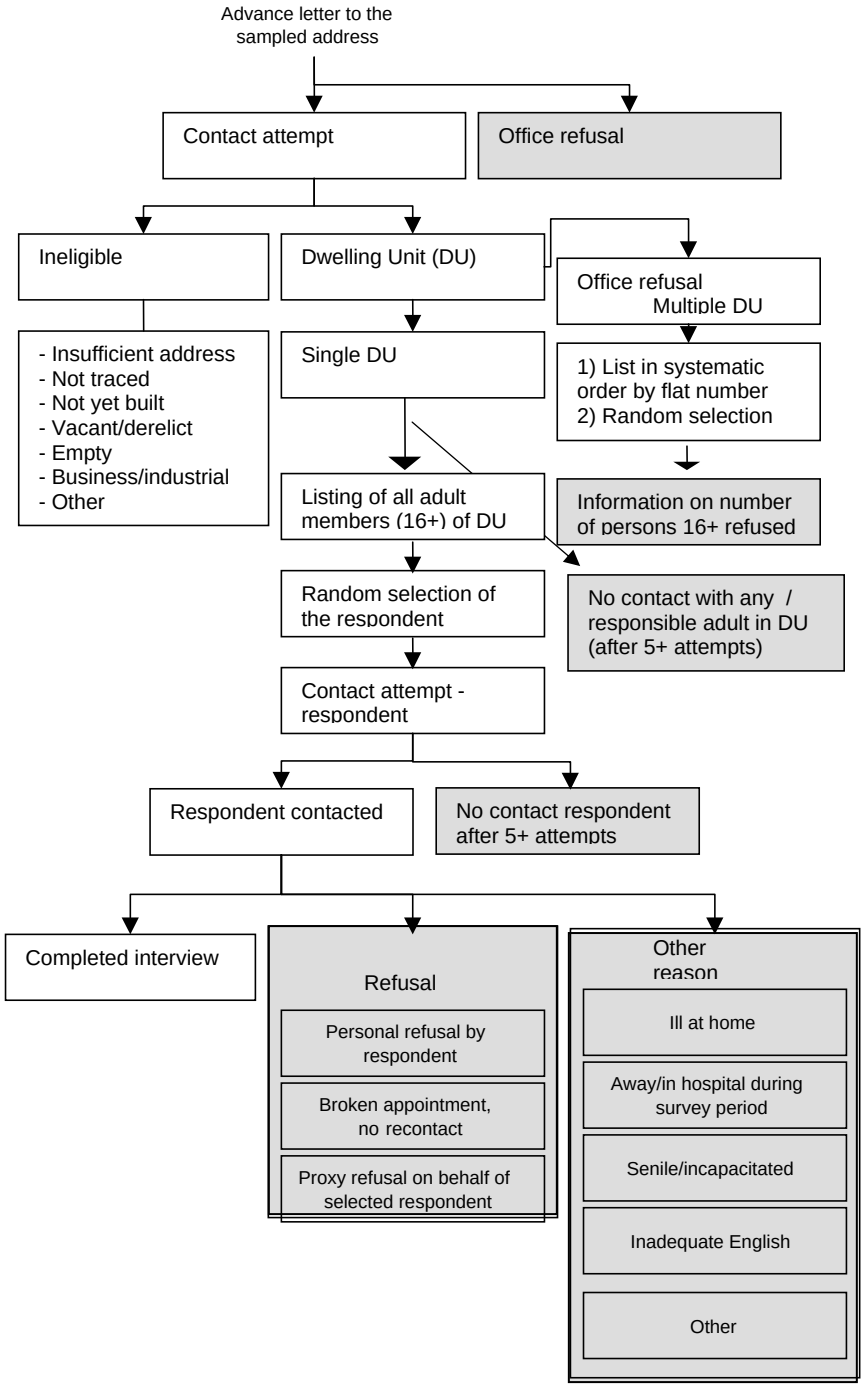

Figure 3.1: The process for a sample of addresses. (cf. Laiho and Lynn (1999).

At all remaining sample addresses, an interviewer attempts to visit the address and make contact with the residents. In the majority of cases, the address will consist of a single dwelling (a house or a flat), in which case the interviewer's task is to list all adult residents and make a random selection of one to interview. Some people refuse to provide the information necessary to list the residents; other people will never be at home when the interviewer visits, resulting in a noncontact. In the small minority of cases where an address contains multiple dwellings, the interviewer has the additional task of selecting one or more dwellings. Once the random selection of a person to interview has taken place, the interviewer must attempt to speak to that person. It may not be the person who provided the information to make the listing, and the selected person may not even be at home, so the interviewer may have to make subsequent visits to the address to find this person. If contact is successfully made, there are still several reasons why an interview may not be achieved. The selected person may refuse, or somebody else may refuse on their behalf (for example, a husband who does not allow the interviewer to speak to his wife, or a parent who does not allow contact with their child—a *proxy refusal*). The selected person may be unable to participate due to illness or incapacity or may not speak adequately the language in which interviews are being conducted. On United Kingdom surveys of this kind, it is often found that around 3% to 6% of sample addresses will result in a noncontact, between 15% and 35% will be a refusal and around 1% to 2% will be a nonresponse for some other reason.

It can be seen that the survey participation process is quite complicated and there are many stages in the process at which there is an opportunity for nonresponse to occur. In general, the more complicated and demanding the process of collecting data is, the more likely it is that nonresponse will occur.

3.3.2 Telephone Surveys of Named Persons

Many surveys are carried out by telephone. In some countries, this is a common method of carrying out surveys of the general population. This usually involves selecting a random sample of phone numbers by a method such as random digit dialling (RDD). Telephone surveys are also often used when the sample is of named persons for whom a telephone number is available, perhaps from the sampling frame or having been collecting in an earlier survey interview. With such surveys, noncontact can occur if the telephone number is incorrect or if the sample member has changed telephone number recently (for example, due to moving home). In some such cases, it will be possible to obtain the new phone number, but not always. If the phone number is correct, noncontacts will occur if the sample member is never at home when the interviewer calls, or if they do not answer the phone. It is increasingly common in some countries for people to use devices that enable them to see the phone number of the person calling them before they answer the phone. They may choose not to answer if they do not recognize the number. And even if contact is made, the sample member may refuse to carry out the interview. It is much easier to refuse on the phone than to an interviewer standing at the door, so it is a big challenge for telephone interviewers to prevent this from happening.

3.3.3 Postal Surveys

Surveys that use self-completion questionnaires administered by post (mail) may seem to be rather simple in terms of the participation process. Either you receive the completed questionnaire or you don't. But in reality the underlying process is still quite complex. The difference is that it is hidden from the view of the survey researcher to a greater extent than with interview surveys. First, there will be some cases where the questionnaire does not reach the intended recipient, because the address is wrong, because of a failure of the postal service, or because someone else at the address intercepts it. Amongst cases where the questionnaire successfully reaches the sample member, there will be several reasons for it not being returned. In some cases this represents a refusal, in the sense that the recipient consciously decides not to complete the questionnaire (but only in a small minority of such cases will the recipient inform the survey organization of this decision), in other cases it may simply be a result of forgetting, as the recipient puts the questionnaire to one side with an intention to complete it later, but then fails to do so. There may be some cases where the respondent is unable to complete the questionnaire due to illness, illiteracy, or inability to read the language of the questionnaire. And some questionnaires may be completed but get lost in the post.

3.3.4 Web Surveys

The nature of nonresponse on web surveys depends heavily on the design of the survey. For invitation-only surveys, where a preselected sample of persons is sent (typically by email) an invitation to complete the questionnaire, noncontact can be considerable. This can be caused by incorrect or out-of-date email addresses, by the recipient's email system judging the email to be spam and therefore not delivering it, or by the recipient judging the email to be spam and not opening it. For web surveys, levels of break-off are typically higher than with other survey modes. This is where a respondent gets a certain way through the questionnaire and then decides not to continue. There are many reasons why this happens and, although the proportion of break-offs can be reduced by good design, it is a considerable challenge. Further discussion of the sources of non-response and what to report can be found on the website of EFAMRO (www.efamro.org), see also de Leeuw, Chapter 7.

3.3.4 Flow Samples

Many surveys involve sampling and collecting data simultaneously from a mobile population that is defined by time and location. Examples include international passenger surveys that sample and interview at ports and airports, surveys of train or bus passengers, and surveys of visitors to a particular location or service such as a national park, a museum, or an employment agency. With this kind of survey, noncontacts are likely to consist solely of cases where the sample person could not be approached as there was no interviewer available to do so. This tends to happen during periods of high flow,

as interviewers are still occupied interviewing previously sampled person(s). The extent to which this happens depends on the frequency with which people are sampled at each sample location (determined by the population flow and the sampling interval) and the number of interviewers working at that location. The extent of refusals will largely depend on the time that sample members have available and the circumstances. If you are attempting to interview people while they are waiting in a queue you may get rather low levels of refusal as the sample members do not have many alternative ways to spend the time. But if you are sampling people who have just disembarked from a train, sample members tend to be keen to continue their journey and refusal levels will be higher.

3.3.5 Business Surveys

Surveys of businesses are different from surveys of households in two important ways that affect nonresponse. First, respondents are not answering on their own behalf but on behalf of the business. This raises a different set of concerns regarding confidentiality and sensitivity of responses, which could affect refusals. Second, it is often necessary for more than one person in the business to contribute to the survey answers and the survey organization rarely knows the identity of these people in advance. Consequently, a response will only be obtained if all the necessary people are identified and contacted during fieldwork. Many business surveys are conducted as self-completion surveys, so this often requires a questionnaire to be passed around the business to each relevant person. The ways in which the survey organization controls and facilitates that process are likely to influence the extent of nonresponse due to a failure to reach the relevant person(s)—a form of noncontact.

3.4 WHY DO PEOPLE REFUSE TO PARTICIPATE IN SURVEYS?

Refusals often constitute a large proportion of survey nonresponse. Consequently, they warrant careful attention. A conceptual framework for survey co-operation in the case of interview surveys is presented in Figure 3.2. The decision about whether or not to co-operate is an outcome of the interaction between interviewer and sample member. The behavior and performance of both the sample member and the interviewer during the interaction will be largely influenced by two sets of factors. These can be broadly labeled the social environment and the survey design. (Both actors in this interaction will of course also have their own personal characteristics and predispositions upon which these two sets of factors act.)

The social environment includes the degree of social cohesion, the legitimacy of institutions, and so on. These influence the degree of social responsibility felt by a sample person and the persuasion strategies and decision-making strategies used by interviewers and respondents respectively. Also, the immediate environment in which the survey interview is to take place

is likely to affect a sample member's willingness to be interviewed. Relevant factors include comfort and perceived safety.

Many aspects of survey design affect response rates. These are discussed in section 3.5 later. Other, broad, aspects of survey design can be considered as constraints upon the interaction between sample member and interviewer. Mode of interview is very important. Interviewers are much more limited in the ways they can communicate with a sample member if they are talking on the telephone rather than standing in front of them face-to-face. They cannot show the sample member documents or identity cards, they cannot use body language or gestures, and so on. These limitations may contribute to the lower levels of success that interviewers seem to have in avoiding refusals on telephone surveys. How interviewers introduce the survey is also likely to be influenced by the length and content of the interview. For example, if a sample member seems generally willing but appears not to have much time available currently, then faced with a long interview an interviewer may suggest that she returns at a more convenient time ("retreat and return") rather than asking to start the interview immediately. But if the interview is short, she may be more likely to suggest starting the interview immediately. These tactics may have different implications for the survey outcome.

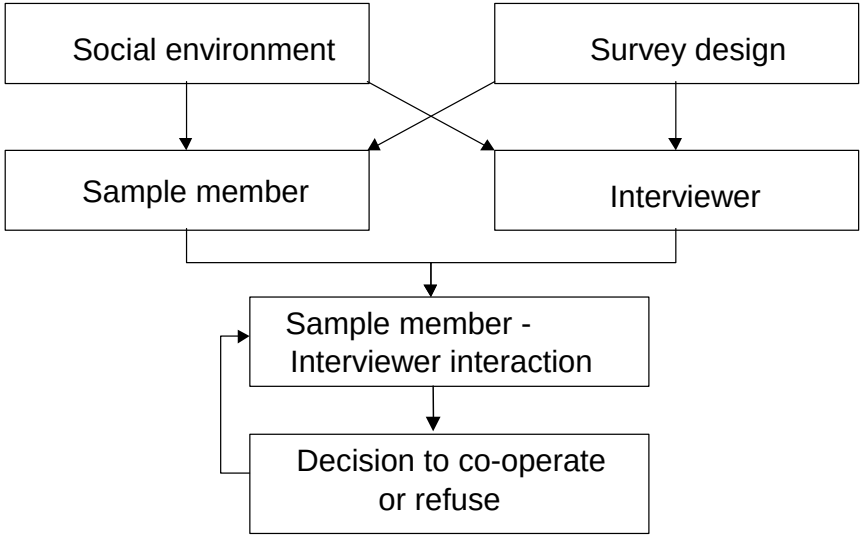

Figure 3.2: A conceptual framework for survey co-operation.
Adapted from Groves and Couper (1998, p. 30).

Groves, Cialdini and Couper (1992) discuss six psychological principles that apply to requests to take part in surveys: reciprocation, authority, consistency, scarcity, social validation and liking. Additionally, three types of attributes of the interviewer may have an important influence on the interaction with the sample member. The interviewer's expectations regarding the likelihood of

gaining co-operation is affected by previous experiences but can also be influenced by appropriate training. Their appearance and manner influence sample members' impressions of the interviewer's intentions and whether it is likely to be safe or desirable to talk to them. The more, and more diverse, previous survey experience the interviewer has had, the more likely it is that they will be able to react to particular situations in appropriate ways that will minimize their chances of getting a refusal.

Survey topic influences some sample members' willingness to respond. The more relevant the survey appears, the more likely sample members will agree to be interviewed. But being interviewed can also have negative consequences, often referred to as the burden of taking part in a survey. For many people, the main component of burden is simply the amount of time that it takes. Other aspects of burden include cognitive effort, sensitivity and risk. Cognitive effort essentially relates to how difficult the questions are to answer. Sensitivity refers to embarrassment, stress or pain that may be caused by the questions. Risk acknowledges that being interviewed may (be perceived to) involve a risk to one's personal safety by letting a stranger into one's home, but also that answering questions that may reveal illegal or immoral behavior could result in being punished for that behavior (or at least be perceived to risk such an outcome).

Ultimately, the sample member must rapidly consider the potential benefits and potential drawbacks of agreeing to the interview and make a decision. The benefits and drawbacks will be weighed up against one another and if the drawbacks appear to weigh more heavily, the sample member will refuse. This idea is nicely encapsulated in the leverage-saliency theory of survey participation (Groves, Singer & Corning, 2000). The survey researcher should therefore, through the behavior of the interviewer and the design of survey documents and materials, emphasize to sample members the benefits of taking part and to de-emphasize the disadvantages. Of course, the various considerations will not be equally important to all sample members and that is why interviewers should be able to tailor their approaches (Groves & Couper, 1998, pp. 248-249) to react to the particular circumstances and concerns of each sample member. Various materials are available to assist in training interviewers in techniques to maximize response rates. These include a video with an accompanying trainers' booklet (National Centre for Social Research, 1999) and an earlier book (Morton-Williams, 1993).

3.4.1 Self-completion Surveys

Tailoring is an important tool to reduce the chance of getting a refusal. However, compared with tailoring by interviewers during an introductory conversation, it is much more difficult to tailor documents such as advance letters, as typically little is known in advance about the sample members or their concerns. This is perhaps one reason why self-completion surveys, when not introduced by an interviewer, tend to achieve lower response rates than interview surveys. The framework presented in Figure 3.2 can be applied also to self-completion surveys, simply by replacing interviewer with survey organization in each box. The interaction with the sample member now

typically consists of the sample member reading written material. In the case of a postal survey, this will be a letter, a questionnaire, possibly one or more reminder letters, and possibly a survey website. In the case of a web survey, the written material comes in the form of an invitation email or letter plus instructions that accompany the questionnaire on the website. The interaction is therefore much more limited and the survey organization rarely has the opportunity to react to particular concerns or circumstances of sample members. Strategies that can be adopted to minimize refusals on self-completion surveys are discussed in Dillman (2000).

3.5 CALCULATING AND PRESENTING RESPONSE RATES

Response rate is an important indicator of the success of the survey at representing the population of interest (assuming the sample was selected by an appropriate probability method). It can also be used as an indicator of the success of the data collection operation. In fact, response rates and other kinds of outcome rates such as eligibility rates, contact rates and refusal rates provide useful information for many purposes. Consequently, the way they are calculated and presented is important (Lynn, Beerten, Laiho, & Martin, 2002).

Every survey should document the outcome rates achieved. These rates should be calculated in clearly specified ways, so that readers can understand exactly which kinds of units have been included in the numerator and which in the denominator of each rate. Ideally, the method of calculating response rate should be consistent with other similar surveys. Some guidance on how to do this appears in AAPOR (2005) and Lynn, Beerten, Laiho, and Martin (2001); for Internet surveys see EFAMRO. Published response rates are often accepted uncritically, but this is misguided as the rate can be sensitive to the method of calculation. This can make comparisons of published response rates fairly meaningless. It is good practice to publish the number of sample cases in each outcome category (e.g., the kinds of categories in Figure 3.1 mentioned earlier) so that users can calculate whichever rates they wish for themselves. We saw earlier in this chapter that there are many possible ways in which nonresponse can arise on a survey. If we want to learn how to improve response rates next time, it is essential to know how prevalent each reason for nonresponse was. A single response rate does not convey that information—a complete distribution of outcomes is needed.

Even more fundamental is the way in which the outcome categories themselves are defined. This too should be documented explicitly. The guidelines referred to earlier provide a set of standard definitions of outcome categories that can be applied to most surveys.

3.6 MINIMIZING NONRESPONSE

A consequence of the diversity of ways in which nonresponse arises is that we need a range of techniques and tactics to prevent nonresponse. No single

technique is likely to have a large impact on response rate. We need to combine many techniques, applied to different stages of the design and implementation process. The classification in Table 3.2 can serve as a useful starting point for thinking about what we should do.

3.6.1 Identifying/Locating Sample Units

Success at identifying or locating sample units largely depends on the quality of information on the sampling frame. Sometimes, it may be possible to augment sampling frame information by matching sample units to other data bases or sources of information. The researcher should consider at an early stage whether this is likely to be necessary and, if so, to set up systems in advance of field work. During field work, it may be appropriate to have systems for locating new contact details for sample members who have moved. This may require interviewers to travel to different areas. Again, such systems require advance planning.

3.6.2 Making Contact

Often, considerable efforts are needed to make contact with sample members. This is particularly true for face-to-face and telephone interview surveys. The necessary extent of the efforts, and the best way to make them, depends on the nature of the sample units and the nature of the survey task. The researcher should consider carefully how, when and where the sample members are most likely to be available to be contacted and to develop field work procedures appropriately. I outline below some techniques that have been found to work well in some common survey situations, but you must think critically about the extent to which these findings are relevant to your survey.

In some countries, particularly industrialized ones, the amount of time that people spend in their home has been decreasing in recent years. Some population subgroups—for example, young single professionals living in big cities—spend very little time at home. This presents challenges for at-home interview surveys. Interviewers can reduce noncontact rates by making more call attempts and by varying the times of day and days of the week of their call attempts. Both of these dimensions of interviewers' calling patterns (number of calls and time/day of calls) are important. In the case of face-to-face surveys, many survey organizations stipulate that an interviewer must visit an address at least 4 (or 5) times, including at least once on a weekday evening and at least once at the weekend, before it can be classified as a noncontact. Often, considerably more attempts are made. With a clustered sample (see Lohr, Chapter 6), each time an interviewer visits the sample area, he or she can make a further call at each address where contact has not yet been made. With a more dispersed sample, the noncontact rate is likely to be higher unless special measures are taken. It is important to provide interviewers with motivation to make extra calls, especially at evenings and weekends. This can partly be achieved by good training, but financial reward will also be needed. Paying a fixed hourly rate provides no incentive for interviewers to call at times when

people are more likely to be at home rather than times when they themselves prefer to work. Paying a modest bonus for achieving a target contact rate could be effective. All these counter measures are, unfortunately, likely to increase the costs of fieldwork and the length of the data collection period.

The marginal cost of making extra call attempts is relatively low on a telephone survey so many attempts can be made. It is not uncommon for survey organizations to stipulate that a sample telephone number must be attempted at least 12 or 15 times before it can be classified as a noncontact. If sample members are being telephoned at their homes, it will be important, as with face-to-face interviewing, for interviewers to work evenings and weekends. As some people can be away from home for long periods (on holiday, on business, etc.), contact rates will be higher the longer the fieldwork period.

If contact is made with someone other than the sample member, it is important to obtain and record information about when the sample member is likely to be available, and subsequently to phone again at that time. This requires a carefully planned call scheduling system. The system should ensure that an interviewer (it may not necessarily be the same interviewer) calls back at an appropriate time if an appointment is made or if an indication is given of when the sample member is likely to be available. Even if no contact at all is made, the call scheduling system should aim to ensure that future calls are made at different times and on different days to the previous unsuccessful calls. On a modest sized survey, the interviewers may do the scheduling using paper based diary systems. On a larger survey, it may be more efficient for a supervisor (perhaps themselves a senior interviewer) to do the scheduling using a spreadsheet or other computer based system. If the work is being carried out from a telephone unit or other central office location, this is particularly likely to be the best solution. Many survey organizations use computer assisted systems for telephone surveys, and these incorporate automatic call scheduling facilities.

If self-completion questionnaires are to be posted to sample members, contact will only be made if the sample member actually receives the mailing, opens the envelope and looks at the contents. The most important determinant of noncontact rate is therefore likely to be the quality of the address information used for the mailings. Once the mailing has arrived at the correct address, the sample member must be motivated to open it. A plain envelope may be best, to avoid it looking like junk mail. The design of postal survey packages is discussed by de Leeuw in Chapter 13.

On web surveys, to make contact typically requires both that a valid email address is available for each sample member (i.e., one that relates to an account that the sample member checks regularly) and that the recipient is motivated to open the invitation email and read it. The subject line of the message and the ‘ender are therefore important. For further discussion of making contact on web surveys, see Lozar-Manfreda and Vehovar (Chapter 14).

Surveys that aim to sample from a flow (as described earlier) are rather different from other surveys in terms of strategies to minimize noncontacts. The important thing is to ensure that field workers are able to deal adequately with periods of high flow. The appropriate strategy depends on the rate of flow, how well the flow can be predicted in advance, and the time taken for field workers to hand out each questionnaire or administer each interview. It may involve having

different numbers of field workers in each sample location, or at different times of day, or using different sampling fractions at different times.

3.6.3 Obtaining Cooperation

To minimize refusals, the survey researcher should: (a) increase (and emphasize) the benefits of taking part, (b) reduce (and de-emphasize) the drawbacks, and (c) address legitimate concerns of sample members.

The survey should be introduced in a way that makes participation seem likely to be interesting and enjoyable. Emphasize the aspects of the interview that people are more likely to find interesting. Explain that the survey serves useful purposes. Provision of payment or a small gift can also help. There is considerable experimental evidence that such incentives can reduce survey refusal rates, though the extent of the reduction depends on the nature of the incentive, the study population and other features of the survey. Offering survey respondents a token of our appreciation helps to establish the *bona fide* nature of the survey and makes them feel better disposed to reciprocate by offering their co-operation in return; however, providing an incentive to each respondent raises costs and survey funders may need to be convinced that it is likely to be cost effective.

For many people, the main drawback of taking part in a survey is the amount of their time that it will take. This should be minimized by keeping questionnaires as short as possible – ask only questions that are necessary; do not ask an open ended question (which might take a minute or two) if a closed question (taking a few seconds) provides equivalent information. People might be more willing to take part at certain times than others. Be flexible and allow them to take part when it is most convenient for them. On interview surveys, the interviewer should be prepared, when it is clear that she has called at an awkward time, to call back later when it is more convenient for the sample member. Otherwise, there is a high risk that a refusal will result. Offer to make an appointment. Some sample members may think that taking part will be too difficult for them, or that the survey is not relevant to them. Tell them that the questions are not difficult and that no specialist knowledge is required. Tell them that you are interested in the views and experiences of *all* kinds of people—that the survey results must represent everyone, not just the people with strong views or expert knowledge.

Sample members may be concerned that their answers should not become known to anyone else. Tell them that the survey is confidential and that nobody outside the research team will be able to link their answers to their name or address (you must, of course, have systems in place to ensure this). Explain that results will be made available only in the form of statistical summaries—no individuals will be identified. Tell them that they will not receive any direct mail as a result of taking part and that they will not be asked to take part in any further surveys (if this is true). On an in-home interview survey, sample members—especially older people—may be reluctant to invite a stranger into their home. Be sure that interviewers carry identification and that sample members are given the name and telephone number of someone who can verify that the survey is genuine. It is good practice to notify the local police station in areas where you are carrying out in-home interviews. Interviewers can tell wary respondents that

the police know about the survey and suggest that they contact the police station to check this if they wish. Interviewers should be prepared to offer to come back when there will be someone else there too, if a sample member is reluctant to let them in while they are alone.

The method of communicating all these messages to sample members depends on the survey. On interview surveys, you will be heavily dependent on the interviewers to explain the survey and answer questions. It is therefore important that interviewers are well trained in what to say to avoid getting a refusal. Depending on the nature of your sample, you may also be able to send an advance letter to sample members. If the letter has an official letterhead, that helps to establish the credibility of the survey. The letter should also provide the name and phone number of someone to whom queries can be directed. (This person, of course, must also be trained in refusal avoidance techniques and must be provided with information necessary to answer most of the sorts of queries and concerns that sample members are likely to raise). The letter should also briefly outline the nature of the survey and explain that answers will be treated confidentially. It should explain that an interviewer will be in touch shortly. It is generally best to avoid mentioning how long the interview will take in the advance letter—leave this to the interviewer to explain.

On a postal survey, the survey documents must convey all the important messages to sample members. Typically, the documents consist of a covering letter and the questionnaire itself. You may also include a leaflet containing further information about the survey or about the organization for whom the survey is being carried out. Sample members will decide, based upon their perceptions of these documents alone, whether or not to take part. Similarly, for web surveys the respondent's perception of the information presented on screen determines whether or not they decide to proceed with the survey.

3.6.4 Minimizing Other Reasons for Nonresponse

To reduce the number of interviews that are lost due to the sample member being too ill or temporarily away, a compromise solution can be to accept a proxy interview from a spouse or other household member, answering on behalf of the sample member. This can sometimes be appropriate, depending on the nature of the survey questions. There is no point asking a proxy respondent about things that they do not know. And it is certainly not possible to ask opinions or attitudes by proxy. In general, if you choose to accept proxy interviews in certain circumstances, there is likely to be a trade-off between response rate and measurement error. Other ways of reducing the number of temporarily absent sample members include extending the field work period and offering alternative modes of response, although these may have other disadvantages.

For many surveys, people who do not speak (in the case of an interview survey) or read and write (in the case of a self-completion survey) the main language (or one of the main languages) of the country are an important subgroup. Excluding them would certainly introduce nonresponse bias. But including them is likely to be expensive. It is necessary to provide translated

materials and, in the case of an interview survey, trained interviewers who speak each language. And translation of survey materials is not a simple matter (see Harkness, Chapter 4), so the translation process must be a careful one.

3.7 NONRESPONSE ERROR

Ultimately, nonresponse is important because it affects estimates. In our earlier example, nonresponse caused us to estimate that 36% of people had low literacy ability when the true figure in the population was 43%. In general, nonresponse introduces error to our estimates if the nonrespondents differ from the respondents in terms of the things we are trying to measure (unless we can fully correct for these differences at the analysis stage—see section 3.8). Suppose we want to estimate a characteristic *Y*. This could be any kind of population parameter: a mean, a proportion, a measure of association, and so forth. We estimate *Y* by the corresponding sample statistic *y*. But we only observe *y* for the respondents in the sample, so the value we observe might differ from the value we would have observed if we had complete response. We can express this as follows:

$$y_r = y_n + \frac{nr}{n}\left(y_r - y_{nr}\right), \tag{3.1}$$

where n is the (selected) sample size; there are r respondents and nr nonrespondents (so $r + nr = n$); y_r is the value of y for the respondents (observed); y_{nr} is the value of y for the nonrespondents (not observed); and y_n is the value of y for the complete sample (not observed).

The amount by which the estimate y_r differs from y_n is the nonresponse error. This is the product of two components. The first, nr/n, is the nonresponse rate. The second, $\left(y_r - y_{nr}\right)$, is the difference between respondents and nonrespondents in our variable of interest. We therefore need to pay attention to *both* these components. The nonresponse error or bias is given by

$$y_r - y_n = \frac{nr}{n}\left(y_r - y_{nr}\right). \tag{3.2}$$

Note that knowledge of the response rate alone does not tell us anything about nonresponse error. It is possible to have a high response rate (small nr/n) but have large nonresponse error (if $\left(y_r - y_{nr}\right)$ is large); it is also possible to have a low response rate (large nr/n) but have little or no nonresponse error (if $\left(y_r - y_{nr}\right)$ is small). To estimate the extent of nonresponse error, we need to find a way to estimate $\left(y_r - y_{nr}\right)$ (see section 3.7). And to minimize nonresponse error we need to minimize *both* nr/n and $\left(y_r - y_{nr}\right)$. The

previous section discussed how we can minimize nr/n, but minimizing $(y_r - y_{nr})$ can be more challenging. Essentially, we need to concentrate on increasing response rates amongst the sample groups who would otherwise be unlikely to respond.

To illustrate the use of this expression for nonresponse error, we return to our literacy example (Table 1). We have $y_r = 180/500 = 0.36$ and $y_n = 300/700 = 0.43$; the nonresponse error $y_r - y_n = -0.07$ is based on $(y_r - y_{nr}) = (0.36 - (120/200)) = -0.24$ and $nr/n = 200/700 = 0.286$, alternatively calculated as $0.286 \times (-0.24) = -0.07$.

3.8 ESTIMATING NONRESPONSE ERROR

Estimating $(y_r - y_{nr})$ is a big challenge as y_{nr} is, by definition, not observed. But there are several possible approaches. Often, more than one of them is possible. It is a good idea to look at every available source of information about nonresponse as this helps you to build up a picture of the nature of nonresponse on your survey.

3.8.1 Use sampling frame information

Many sampling frames are a useful source of auxiliary information about each unit. If we include this information on the sample file, we can use it to compare respondents and nonrespondents.

Table 3.3: Estimating nonresponse error using sampling frame data

Highest qualification	Response rate	Selected sample %	Responding sample %
1. 5+ Higher grades	91.1%	18.0	21.4
2. 3-4 Higher grades	85.1%	13.0	14.5
3. 1-2 Higher grades	81.7%	15.0	16.1
4. 5+ Standard grades 1-3	76.4%	8.1	8.1
5. 3-4 Standard grades 1-3	74.1%	9.1	8.8
6. 1-2 Standard grades 1-3	69.1%	14.5	13.1
7. Standard grades 4-7 only	62.6%	14.4	11.8
8. No qualifications	59.6%	7.8	6.1
N		4,542	3,469

Source: Lynn (1996)

Table 3.3 presents an example, using data from the Scottish School Leavers Survey, a postal self-completion survey of young people aged 16 to 18 in

Scotland. The sampling frame for this survey includes a record of examination passes achieved at school. This information has been used to derive an ordinal variable with eight categories, shown as rows in Table 3.3.

Because we know the level of qualification achieved by each sample member, whether or not they responded to the survey, we can calculate response rates separately for each group. The response rate is highest amongst the most highly qualified sample members (91.1%) and lowest amongst those who left school with no qualifications (59.6%). Thus, we can obtain a direct measure of nonresponse error in, say, the percentage of people leaving school with very low qualifications: $y_r - y_n = 17.9 - 22.2 = -4.3$. However, it is not immediately helpful to know that nonresponse would cause us to underestimate this percentage by 4.3 if we used the responding sample, because we already know the percentage for the complete sample. The usefulness of the statistic lies in the fact that leaving school with very low qualifications is correlated with other parameters that we might wish to estimate using the survey data, such as labour market outcomes. We could be fairly sure that nonresponse error would cause us to underestimate the proportion of young people who are unemployed at age 20, for example, although we would not know by how much. Using sampling frame data thus has the advantage that nonresponse error can be calculated directly, but the disadvantage that this can only be done for the auxiliary variables and not for survey variables. Typically, it requires advance planning as we need to capture the auxiliary data during the process of sample selection.

3.8.2 Using Linked Data

It may be possible to link data from other sources to the sample records (see Bethlehem, Chapter 26). Only rarely is this possible for individuals, as in most contexts this requires the individuals' consent (which cannot be obtained for nonrespondents). But linkage is often possible at some higher level of aggregation. For example, in many countries a range of population statistics are published for small areas, either from a Census or from administrative data (e.g., on zip code level). The sample for a general population survey can be linked to such auxiliary data provided that suitable geographic identifiers exist on the sample file. The data can then be used in the same way as for sampling frame data.

3.8.3 Interviewer Observation

For an in-home face-to-face interview survey (and some other types of survey) it can be possible to ask interviewers to record certain characteristics of each sample unit from observation. For example, this might include the type of dwelling, the construction materials, the age of the dwelling, the nature of the surrounding area, and so on (e.g., Lynn, 2003b). The data on these characteristics can then be used in the same way as for sampling frame or linked data. A variation on interviewer observation is to collect data about nonrespondents by proxy, for example from neighbors or work colleagues. This

is rarely very satisfactory as a means of studying nonresponse, as the data are typically far from complete and it cannot be assumed that measures are comparable with those collected from the respondents themselves.

3.8.4 Comparison with External Data

Sometimes there exist aggregate data about the population under study from some external source such as a recent Census or administrative data. If these data relate to one or more of the same variables about which data have been collected by the survey, then the responding sample can be compared with the population data; however, there are two important things to note about such comparisons. First, any differences between the two sources may not be due (solely) to nonresponse. Other factors affecting the comparison include coverage error and sampling error. These factors are confounded. Second, the data themselves may not be strictly comparable. There may be differences in the time period to which they refer, in the reference population to which they relate, and in the way they have been collected. Some data items may be more sensitive than others to such differences. In consequence, some observed differences between the responding sample and the external data may not reflect any real difference at all—rather, they may simply be due to differences in the way the variables have been measured. If you are planning an external comparison, consider carefully which variables are likely to be least sensitive to differences in the way the data were collected.

3.7.5 Using Process Data

Often, survey researchers can learn a lot from information about the process of collecting the survey data. For example, for an in-home survey, it is possible to record the number, timing, and outcome of all visits made to each sample unit before the interview was achieved; for a telephone survey you can record the number, timing, and outcome of all calls; for a postal survey you can record the number of days until the questionnaire was received or the number of reminder mailings that had to be sent to each unit. Process data of this kind, also often referred to as para data (see also Mohler et al, Chapter 21), can be available for all sample units. You can then observe how these data relate to the survey variables to obtain an indication of the likely direction and magnitude of nonresponse bias.

3.8.6 Survey of Nonrespondents

After a survey is complete, a sample of the nonrespondents can be selected for intensive follow up. This can be enlightening, but it is very hard to get a good response rate to a survey of nonrespondents. Ultimately, the follow up survey only tells us something about the relatively more accessible and less unwilling nonrespondents and we will not know how representative they are of all nonrespondents. In short, this survey too suffers from nonresponse error.

3.8.7 Panel Dropouts

In the case of panel surveys and other follow up surveys, we are in a strong position to understand the nature of nonresponse subsequent to the first wave. For the first wave, we still have to use one or more of the methods described earlier. But for subsequent waves, we can use all of the survey data collected at the first wave, and any other wave prior to the one being studied, as auxiliary data. The advantage of this is that we typically have a rich range of variables available and at least some of them are likely to be highly correlated with the survey variables of interest. Often, they are measures of exactly the same concept, relating to an earlier point in time.

3.9 ADJUSTMENT FOR NONRESPONSE

Understanding something about the nature of nonresponse and the likely impact of nonresponse error on survey estimates is important. But rather than simply describing it, it is better to adjust the estimates for it. This can be done quite simply using weighting. However, although it is simple to implement nonresponse weighting, it is not necessarily so easy to identify a *good* way of weighting amongst the possible ways that present themselves. Care is needed.

Consider again the data of Table 3.3. The response rate amongst sample members in category 1 was 91.1%. If we give each respondent in category 1 a weight of 100/91.1 (i.e. 1.098) in our analysis, and applied a similarly constructed weight to respondents in each of the other seven categories, then the categories would be represented in their correct (selected sample) proportions in the analysis. This makes intuitive sense, as every 91.1 respondents in category are in some sense representing 100 selected sample members, so they must be given extra weight to represent the additional missing 8.9 sample members. The weights will be greater the lower the response rate: in our example the largest weight is 1.678 for respondents in category 8.

After weighting has been applied, the nonresponse error that remains in a weighted estimator can be expressed as follows:

$$y_{rw} - y_n = \frac{1}{n}\sum_{h=1}^{H} nr_h \left(y_{r_h} - y_{nr_h} \right)$$

where there are H weighting classes, denoted $h = 1, \ldots, H$ ($H = 8$ in our example).

It can be seen that the error is now a weighted sum across the weighting classes of the difference in y between respondents and nonrespondents. In other words, the error no longer depends on differences *between* the classes, as this is what the weighting has corrected. The definition of the classes is therefore important. For nonresponse weighting to be successful, four criteria should be met: (a) Response rates should vary over the classes; (b) Values of target variables (y) should vary over the classes; (c) Respondents and nonrespondents should be similar to one another *within* each class (i.e. $y_{r_h} - y_{nr_h}$ should be small); (d) Class sample sizes should not be too small. When choosing between

alternative ways of creating weighting classes, these criteria should provide guidance. Weighting is discussed in more detail by Biemer and Christ in Chapter 17. An important point to remember at this stage is that it will not be possible to implement effective weighting unless you have planned ahead and collected some of the kinds of data outlined in the previous section.

3.10 CONCLUSION

Nonresponse is important and there are many different ways in which it can arise. Equally importantly, there are many different things that we as survey researchers can do to combat the undesirable consequences of nonresponse. Almost every stage of the survey design and implementation process has the potential to affect nonresponse error. Consequently, we must keep the issue of nonresponse in mind at all times. When specifying the sample selection method, we should consider whether there are useful data that can be captured from the sampling frame and that will help us later with nonresponse analysis and possibly weighting. When designing field control documents and sample control systems, we should consider whether there are useful data that can be collected by interviewer observation or as indicators of the difficulty of obtaining a response from each unit. When recruiting and training interviewers, we should place an emphasis on the kind of social skills needed to avoid refusals and on working patterns that will minimize noncontacts. Data collection procedures should incorporate appropriate reminders or multiple attempts to contact sample members. Questionnaires should be attractive, interesting, and not too demanding or intrusive. And so on. There are many things we can do to minimize the impact of nonresponse and there are many success stories of surveys that have successfully improved response by reviewing their procedures and implementing a coherent set of changes.

Nonresponse will therefore be a theme throughout this book. In almost every chapter you will find references to it. Tackling nonresponse involves carrying out every stage of the survey in a thoughtful, careful and thorough manner. In short, good survey practice.

GLOSSARY OF KEY CONCEPTS

Adjustment. A term applied to a number of post fieldwork procedures, such as weighting and imputation, that can be used to reduce nonresponse error.
Noncontact. Failure to communicate with a selected sample unit and to inform the unit of their selection for the survey.
Nonresponse. Failure to obtain useable survey data from an eligible selected sample unit.
Nonresponse error. The difference between a survey estimate and the equivalent estimate that would have been obtained if all selected units had responded.
Refusal. A decision by a selected sample unit not to respond to the survey.

Chapter 4

Comparative Survey Research: Goals and Challenges

Janet A. Harkness
Director Survey Research and Methodology Program,
Director Gallup Research Center,
University of Nebraska-Lincoln, USA
Senior Scientist at ZUMA, Mannheim, Germanyy

4.1 INTRODUCTION

This chapter considers some of the key challenges to achieving comparability in deliberately designed cross-cultural and cross-national surveys. As the word challenge reflects, we focus on topics for which theoretical frameworks or current solutions are less than perfect. We spend some time therefore on issues of standardization and implementation, on question design and on question adaptation and translation. Among the topics not dealt with here, but of obvious relevance for comparative survey research, are sampling, analysis, instrument testing, study documentation, and ethical considerations. See Häder and Gabler (2003), Lynn, Häder, Gabler & Laaksonen (2007), Lepkowski (2005) on sampling in cross-national contexts; Saris (2003a, 2003b), Billiet (2003), van de Vijver (2003), and contributions in Hambleton, Merenda, & Spielberger (2005) cover important issues in instrument testing; on documentation see Mohler, Pennell & Hubbard (Chapter 21) and Mohler and Uher (2003) and on ethical considerations see Singer (Chapter 5).

Because numerous terms used in the chapter are understood in a variety of ways in different disciplines, we explain how these are used here. The term *comparative* is used to refer to any research that is designed to compare populations. The term *cross-cultural* is used to refer to research across cultural groups either within or across countries. *Cross-national* will be used as a general term for research involving more than one country or nation. Throughout the chapter the emphasis is on *multinational* surveys, that is, surveys across multiple countries or nations. In many instances multinational surveys are more complex than within-country cross-cultural research, but they have many basic challenges in common. *Multilingual* surveys are surveys conducted in numerous languages. These can obviously be cross-national studies but may also be national studies. For example, to collect data from multiple immigrant groups, the 2000 US Census was conducted in 6 languages and support was provided for 49 languages (www.facts.com/wusp3006y5.htm). In the Philippines, a country currently reckoned to have about 170 languages, International Social Survey Program (ISSP) modules are fielded using

questionnaires in five languages. In South Africa, ISSP modules are fielded using five written translations and several orally translated versions (see Harkness, Schoebi, Joye, Mahler, Faass, & Behr, 2007, on quality issues in orally translated interviews). Multilingual surveys may or not be comparative with respect to questionnaire design; some may merely be translations of a survey designed for a single context. *Multiregional* surveys collect data at regional levels. The regions may be within-country regions but can also cover regions above the country level, such as southern Mesoamerica (including Nicaragua, Costa Rica, and Panama) versus northern Mesoamerica (covering Belize, Guatemala, and Mexico).

In the course of the chapter we refer to *source* questionnaires or languages and *target* questionnaires or languages. Following usage in the translation sciences, the source language is the language translated out, and the target language is the language translated into. *Questionnaire* is used here to refer to the set of questions that make up a study. This might consist of several sub-sets of questions. In some disciplines these would be called *instruments*, in others, *modules*. In this chapter, however, *instrument* is used as an alternative to *questionnaire*. Distinctions are also possible between *questions* and *items* and between *item scales* and question *batteries*. Thus a Likert-type format of a question might contain multiple statements (the items) that would be assumed to form a scale. Items grouped together for other reasons would simply form a set or battery. Finally, we use the term *general survey research* to refer to research and research methods in which (cross) cultural considerations play no deliberate, active role with regard to design or implementation.

4.2 GROWTH OF MULTINATIONAL, MULTILINGUAL SURVEYS

Into the 1970s, cross-national analyses were still often based on data collected at national level for national purposes that were recoded according to a comparative scheme developed ex post (cf. Gauthier, 2000; Rokkan, 1969). In the intervening decades, deliberately designed cross-national research has burgeoned in every field that uses survey data, with marked growth in the number, size and diversity of studies undertaken, the disciplines involved, the kinds of instruments used, and the cultures and languages accommodated. Twenty years ago, Parameswaran and Yaprak (1987, p35) emphasized the need for better cross-national measurements in consumer research in the face of "explosive growth in the multinationalization of business."

Data collected at national level for national purposes are also still used to make analyses at the supra-national level. Indeed, in developing countries, national data may be all that are available. Comparative uses of national data raise their own particular sets of problems. Mejer (2003), for example, discusses efforts to harmonize social statistics in the European Union; Smid and Hess (2003) discuss challenges related to cross-national market research, and Barnay, Jusot, Rochereau, & Sermet (2005) discuss the problems faced in trying to compare health data across different studies.

Multinational survey data are used both as primary sources of information and in combination with data from other sources such as official statistics, records, and specimens from people, places, or animals. Large-scale surveys and harmonized data studies provide cross-national data for key public domains; education and psychological testing, health, labor statistics, population demographics, and short and longer term economic indicators across multinational regions. In the private sector, data from global marketing studies, consumer surveys, establishment surveys, and media research inform production, planning, and resource allocation.

Changing patterns of immigration have increased cultural diversity in many developed countries and the need to collect accurate and reliable information has resulted in an increase in within-country multilingual research. Sometimes these studies aim to produce national estimates that are as unaffected as possible by bias related to culture and/or language differences. At other times, minority populations are deliberately targeted to gain insight into their living conditions, access to facilities, or family composition. In the coming decades, ensuring adequate language coverage in national surveys may become a pressing issue in some countries, as different linguistic communities do or do not gain high fluency in the country's majority language(s) and as the majority languages possibly cease to be that.

As in national research (cf. Converse and Presser, 1986), questions or questionnaires developed for one context are frequently used elsewhere. Sometimes the goal is to compare findings across studies. In other cases, questions are re-used simply because they have already proved themselves useful. As a result, translated questions may be used verbatim or in translation around the globe. Examples can be found in every discipline: indicators of economic development, of well-being, of product or service satisfaction, of socioeconomic status or human values, as well as medical diagnostic instruments, pain indexes, human skills and competence measurements, and personality assessment are used repeatedly in different contexts and languages throughout the world.

The need for global data has led to a new surge of interest in how best to undertake cross-cultural and cross-national survey research. Similar developments can be noted in the 1940s, in the 1960s and again in the 1980s (cf. for example, Hantrais & Mangen, 1996; Scheuch, 1990; Peschar, 1982; Armer & Grimshaw, 1973; Rokkan & Szczerba-Likiernik, 1968; Rokkan, 1962). Researchers entering the field of general survey research can draw on an array of guidelines, best practice standards, protocols for key procedures, and a rich survey methods literature. Unfortunately, there is not a correspondingly comprehensive and accessible set of tools and guidelines available for multinational survey research. It is therefore not easy for researchers entering comparative research to be sure how best to proceed. In the editors' preface to a book considering qualitative and quantitative research, Hantrais and Mangen note: "Notwithstanding this impressive outburst of research activity, it remains true that few social scientists have been trained to conduct studies that cross national boundaries and compare different cultures" (1996, p. 16).

Can researchers follow best practices as advocated in the general survey context? If so, why do these not always produce the results expected?

Must researchers be informed about the countries, cultures, and languages involved in order to conduct comparative research? What can they do to try to ensure that data collected are valid and reliable? Who can collect the data and how should this be done? Are there informed networks to approach for help? The remaining pages of the chapter address these and other questions.

4.3 TOWARD A COMPARATIVE RESEARCH METHODOLOGY

Discussions of comparative survey research often remark that all social science research is comparative and researchers have often debated whether there was anything particular or different about cross-national research (cf. Lynn, Lyberg & Japec, 2006; Øyen, 1990; Teune, 1990; Lipset, 1986; Grimshaw, 1973).

Acknowledging that social science research is based on comparison does not resolve the question whether different methods are needed for different forms of this research. As Johnson (1998, p. 1) notes: "A major source of the criticism directed at cross-cultural research, in fact, has been the uncritical adaptation of the highly successful techniques developed for monocultural surveys."

Multinational survey research has much in common with other survey research and researchers entering the field should therefore have a solid understanding of general survey research methods and the principles of research in their respective discipline. Nonetheless, we suggest that the methods and the perspectives required for comparative research differ in some respects from those of non-comparative research. In mono-cultural research, for example, questions mirroring the culture, containing culturally tailored language and content and possibly tapping culture-specific concepts, are likely to be the *successful* items. In comparative research, such questions would count as culturally biased and would require to be modified, or accommodated or possibly excluded in the analysis. In non-comparative research, valid and reliable data are critical. In comparative research, data must be valid and reliable for the given national context but must also be comparable across contexts.

At the same time, one can design and analyze comparative research without deciding whether the differences are truly qualitative or not. Grimshaw (1973, p. 4), for example, bridges the divide as follows: "My argument is that while the problems involved are no different in kind from those involved in domestic research, they are of such great magnitude as to constitute an almost qualitative difference for comparative, as compared to non-comparative research."

There *is* general agreement in the literature that multinational research is complex (e.g., Lynn et al, 2006; Øyen, 1990; Kohn, 1987; Grimshaw, 1973; Verba, 1971; Zeldich, 1971). In addition, as Kohn (1987) points out, it is also expensive. Nevertheless, the increased complexity and costs of multinational research are not always matched by an increased sophistication of methods. In fact, the methods adopted in multinational survey research frequently do not reflect more recent developments in general survey methodology. With the

exception of Quality of Life research (cf. Skevington, 2002; Murphy, Schofield & Herrman, 1999), few comparative studies report using cognitive testing, focus group input, expert consultations or extensive pre-testing to develop questions (cf. Smith, 2004). In addition, standards accepted as best practice in survey research at the national level, are often not targeted in multinational research (Harkness, 1999; Johnson, 1998; Jowell, 1998). It may be difficult in the multinational context to find sufficient funding to meet such standards, in that everything has to be paid for multiple times. Many multinational studies certainly do not pre-test *draft* versions of the source questionnaire in multiple countries because of the costs this would incur for translation of questions that might never be used. Translated versions of the *finalized* source questionnaire may be pre-tested, as in the European Social Survey, but such pre-tests are not intended to contribute to source questionnaire development. In addition, as Lynn (2003a) notes, the variability of features in the cross-national context makes it more difficult to set common standards. Documentation of procedures may also be poor (see, for example, Herdman, Fox-Rushbie & Badia, 1997, on translation procedures and their documentation; Harkness, 1999, on quality monitoring; Mohler and Uher, 2003, on general documentation in the comparative context).

4.4 COMPARABILITY AND EQUIVALENCE

In cross-national research, the pursuit of data quality is simultaneously the pursuit of data comparability. Comparability is often discussed in the literature in terms of equivalence. Johnson (1998) counts 52 definitions of equivalence within the social and behavioral sciences. In many instances, functional equivalence, understood as having questions perform in the same way across different populations is targeted through question translation, and numerous kinds of translation equivalence are referred to in survey literature. However, as Snell-Hornby (1988) indicates, the translation sciences also use the term equivalence in multiple ways. In this chapter, when referring to the fact that properties of data, questions, meanings or populations, and so forth admit and justify comparison, we prefer to use the term *comparability.*

Researchers use whatever means are available to try to ensure that data from different populations do permit comparison. A strategy frequently adopted is to keep as much in the project as similar as possible, for example, to ask the same questions, to use the same method of data collection, to standardize interviewing methods with a view to reducing variance in interviewer effects, and to use probability sampling designs. In practice, it is neither possible nor always desirable to implement the same detailed protocols everywhere. For instance, the legal definition of what counts as a refusal and whether refusals can be converted varies from location to location. Properly speaking, anyone declining to participate in Germany is a refusal. Once coded as such, the person should not be re-approached. In other locations, saying no need not immediately count as a final refusal, hence the concept of refusal conversion. The greater restrictions in some locations on interaction with targeted sample units can obviously affect response rates considerably.

4.5 STANDARDIZATION AND STUDY SPECIFICATIONS

The goal of standardization is to enhance comparability; inappropriate standardization may do just the opposite. Appropriate standardization is thus crucial. Because it is neither desirable nor feasible to keep everything the same, study designers have to identify what must be standardized to ensure comparability and at what level this standardization should take place. However, standardization, in particular with respect to data collection procedures and protocols, is an area in which much must still be shared and learned. The following examples illustrate some of the problems.

Some places are inaccessible in winter, others again only properly accessible during winter; Chile is only one example of a country with many climate zones. Thus deciding to standardize fielding periods rigidly can be impractical and disadvantageous. Cultures also differ in the times at which they eat, sleep, work, and so forth. As a result, fixing contact times rigidly across countries would be counterproductive. Thus decisions about the best time, say, to contact sample units must take local conditions into account.

At the same time, awareness of strategies to optimize contact attempts may differ from survey culture to survey culture. It may therefore be important to negotiate minimum contact requirements for every location and to discuss and share tactics known to have worked for other locations. In this way, local conditions can be taken into account and information also shared about strategies that have been used in various contexts. Since procedures that are unfamiliar may at first be declared unsuitable or impractical, it is also important to strike a balance between recognizing local constraints and encouraging local actors to adopt or adapt useful techniques.

A complicating factor in this is that one and the same procedure may produce different effects in different contexts. The Swedish participants in the 2002 European Social Survey (ESS) were convinced they could increase response by making advance telephone contact. French agencies sometimes make the same point. Blohm and Koch (2004), on the other hand, found that advance contact by telephone in the German context reduced the propensity of people to participate. Such findings may reflect cultural differences in norms of communication or in the use of the telephone, or simply reflect interviewer proficiency or preferences.

Decisions about standardization determine the specifications for a study. *Study specifications* are intended to be explicit descriptions of the design and implementation requirements that hold for all participants. They can also specify the means by which different steps are to be achieved (e.g., whether contact can be made by mail, phone, or only in person). Examples of mostly top-down specifications for a European social science study can be found on the ESS web site (European Social Survey site: www.europeansocialsurvey.org).

The challenges involved in implementing decisions and in monitoring compliance with specifications should not be underestimated. Misunderstanding of specifications, or the goal of these specifications, is likely to lead to non-compliance. Intensive discussion of the meaning of specifications and the steps needed to implement them will often be the only route to full understanding.

The desire to excel and to be seen as excelling, often coupled with a lack of expertise in one or more areas, may also encourage non-compliance with required specifications. Here too, we lack a general handbook of shared experience, lessons learned and of "how-to-do-despites."

In top-down designs, external design requirements are fixed first (e.g., face-to-face interviewing) and specifications at national levels articulated later. In a bottom-up approach, conditions at local levels shape the formulation of the general study specifications (e.g., the likelihood of third party presence in interviews determines the design). The most viable mix will often lie somewhere between, with general requirements deciding critical specifications (e.g., that multiple contact attempts are made) although local constraints inform how specific these requirements are and shape the protocols for local elaborations or deviations. Special efforts may be needed to ensure that accurate information about local constraints is collected. Some studies are fortunate enough to be able to finance international meetings of participating teams or visits by information-gathering teams to local sites. Less well-funded projects need to exchange information by other means. Some form of E-conferencing could be useful here. Distributing information collected to all involved can actually stimulate further input. Indeed, some participating units (countries or minorities) may only fully recognize it is appropriate for them to contribute once they see input from other participants. Here too, unfamiliarity can foster uncertainty and rejection, a point to be considered in deciding which specifications are truly viable and which not.

4.6 DESIGNING QUESTIONS

This section describes basic approaches used to design questions in comparative research. At present, we lack an overarching framework for how to apply what we know about question design from general survey research to comparative contexts. The literature on specifics of question design in the comparative context is thus somewhat fragmentary. Moreover, approaches differ depending on the discipline and on the type of instrument involved.

Although he does not address the issue of a general framework, Smith (2003, 2004) provides numerous useful examples and extensive references for individual aspects of questions that may be affected by cultural and linguistic issues, from response scale design, to layout and visual aids, to wording, ambiguity and social desirability. A number of health and education projects also outline their particular models of question design in some detail (e.g., the EORTC Quality of Life guidelines described by Blazeby, Sprangers, Cull, Groenvold, & Bottomley, 2002 and the TIMMS and PISA websites[1]). Harkness, van de VIjver, and Johnson (2003) provide a general overview of question design models that is in part followed in this chapter. Braun and

[1] Trends in International Mathematics and Science Study (TIMSS) site: http://timss.bc.edu;_Programme for International Student Assessment (PISA) site: http://www.pisa.org.

Harkness (2005) discuss the interdependence of meaning and context, indicating how differences in socio-cultural context affect how a respondent perceives what a question means. Culture can determine whether information is considered relevant (cf. Smith, Christofer & McCormick, 2004 for health issues among American Indian women). Schwarz (2003) reports differences across cultures in response to the same response scale stimuli; and Haberstroh, Oyserman, Schwarz, Kühnen, and Ji (2001) illustrate how design modifications can affect what is often assumed to be cultural response behavior. Anderson (1967) and Tanzer (2005) illustrate how comparative design needs to consider visual aspects of instrument design. Authors in Hoffmeyer-Zlotnik and Wolf (2003) and in Hoffmeyer-Zlotnik and Harkness (2005) discuss design and comparability issues for so-called background variables such as income, education, religion, occupation, and race and ethnicity.

Response scales and response styles are more frequently discussed topics. Authors such as Lee, Rancourt, & Särndal (2002) and Chen, Lee, & Stevenson (1995) discuss difficulties encountered in trying to replicate features of Likert-type scales in Asian languages. Ewing et al (2002) discusses four different response scales in the cross-national advertising context; Skevington and Tucker (1999) describe the WHOQOL approach to answer scale development; Johnson, Kulesa, Cho, & Shavitt (2005), Johnson and van de Vijver (2003), Gibbons et al (1999), Baumgartner and Steenkamp (2001), and Javeline (1999) discuss different aspects of social desirability, response styles, and acquiescence in cross-cultural contexts.

Pre-testing is part of questionnaire design refinement. Smith (2004, p. 450f) reviews current practices in cross-national testing and notes that "most cross-national studies fail to devote adequate time and resources to pretesting." When pretesting is conducted, techniques developed in general survey research are applied to instruments intended for cross-cultural implementation. In various places we discussed the interdependence of cultural context and cultural meaning and how this determines whether questions are understood or understood in the same way across cultures. Such cultural differences carry over into discourse. We must therefore be wary about assuming that pragmatic features of discourse are also shared across contexts, assuming, for example, that a sensitive question calling for covert disclosure in context A is sensitive and requires covert disclosure in context B (cf. Kim, 1994, Smith et al, 2004). Recent descriptions of cognitive pretesting, mainly for minority populations in the United States context, are Warnecke & Schwarz (1997), Miller (2003), Willis (2004) and Goerman (2006). Schmidt and Bullinger (2003) point to perceived inequalities in QoL research with regard to within-country testing for minorities. Harkness, van de Vijver, & Johnson (2003) and Harkness and Schoua-Glusberg (1998) outline various techniques used in different disciplines for testing translated questions.

4.6.1 Basic Options for Design

In producing questions for multinational implementation, question design teams have three basic decisions to make. First, they can decide to ask the *same* questions of every population or they can decide to ask *different* questions of

each. A mixed approach based on these choices can combine a set of country-common questions with other country-specific questions. This is sometimes called an emic-etic approach (see 4.6.3). A second and related decision is whether researchers want to *adopt* existing questions (i.e., replicate), *adapt* existing questions (i.e., modify) or, alternatively, *develop new* questions for their study. In many instances, all three strategies may be used in one study. Harkness, van de Vijver, & Johnson (2003) outline the advantages and disadvantages associated with each option: adapting, adopting and writing new questions. Thirdly, researchers also implicitly or explicitly decide on the degree of cross-cultural input they intend to target in their instrument development (see 4.6.2).

Much survey research, comparative or not, is based on using existing questions verbatim for new studies or in modified, adapted form. Questions are often replicated, for example, to compare measurement across time. However, questions may also be modified to accommodate new needs or new contexts. For example, instruments developed for adults can be adapted for children (cf. de Leeuw & Hox, 2004); questions that have become out-dated can be up-dated (Porst and Jers, 2005); or instruments designed for business and commerce can be tailored for use in an academic setting.

In the cross-cultural context, researchers also prefer to use existing questions verbatim or, if this is not possible, in an adapted form. Close translation has traditionally been preferred to more free translation. In each case the assumption is that closely translated questions will succeed in conveying the same stimulus for a new population. Harkness (2003) and Harkness, Pennell, & Schoua-Glusberg (2004) discuss the general challenges of such close translation. As Peschar (1982, p. 65) notes: "However, a literal translation of items and questionnaires does not guarantee the equivalence of instruments…Therefore *functional equivalence* is a much more important objective in comparative research" (emphasis original). Greenfield (1997), Herdman et al (1997), and Herdman, Fox-Rushby, & Badia (1998) are skeptical about how suitable translated survey instruments are for new contexts.

4.6.2 Simultaneous, Parallel and Sequential Approaches

Cross-cultural Quality of Life (QoL) research distinguishes between *sequential*, *parallel* and *simultaneous* approaches to question design. Differences can be found in the way the terms are used and explained in the QoL literature (cf. Skevington 2002; Bullinger, 2004; Anderson , Aronson & Wilkin, 1993; van Widenfelt, Treffers, de Beurs, Siebelink & Koudijs, 2005; the Medical Outcomes Trust Bulletin, 1997) and we do not attempt to resolve these here.

Generally speaking, the terms reflect something about when cultural considerations are considered in questionnaire design, how these are taken into account, and whether the questionnaires in different languages that aim to be functionally equivalent are translations of a source instrument or developed by other means. Simultaneous development targets the highest degree of cross-cultural involvement and sequential development the least. The simultaneous approaches described in QoL literature may aim to have each culture develop

its own questions or to have repeated and considerable cross-cultural discussion of a common set of items. The initial draft items may stem from different cultures and languages. Descriptions of elaborate QoL multi-stage approaches can be found, for example, in Bullinger (2004), Skevington (2002, 2004) and the WHOQOL Group (1994). Parallel designs target cross-cultural input early in the conceptual and question development stages or a common instrument. This is sometimes achieved by collecting items from all the participating countries (cf. Bullinger, 2004) or, as in the ISSP, by having a multi-cultural drafting group develop a set of questions of less varied origin. Sequential models focus on having different populations asked the same questions, with little emphasis at the question development stage on cross-cultural input. Further details are provided later.

4.6.3 Ask-Different-Questions Models

One of the great appeals of asking different questions is that one does not need to translate. Another attractive feature of Ask-Different-Questions (ADQ) models is that the country-specific questions used can relate directly to the issues, terminology and perspectives salient for a given culture and language. A third advantage is that the development of a questionnaire can be undertaken as and when needed. Countries might therefore develop their instruments at the same time (in a sense, simultaneously) or, if joining an existing project at a later date, develop their own country-specific and country-relevant questions as these are required. ADQ approaches are sometimes described as functional equivalence strategies. However, because the questions in any kind of comparative study are required to be functionally equivalent, we have coined the term ADQ. A basic procedure is as follows:

- The design team decides on the concepts and constructs to be investigated and any other design specifications they might make;
- Country– or population-specific questions are designed that collect the locally relevant information for a given construct;
- Versions for different countries and languages can be produced in a collective effort or developed by different teams as the need arises.

A practical example illustrates how ADQ might work and also highlights challenges incipient in the approach. (British) *trousers*, (Scottish) *kilt*, and (Indian) *dhoti* could be considered to be functionally equivalent articles of male apparel, all being coverings for the lower part of the body. Distinctions among them exist nonetheless, such as the contexts in which the garment might be worn (everyday wear vs. festive occasions) or the degree of leg coverage afforded. Such differences might be relevant for some comparisons and irrelevant for others. In similar fashion, the following questions might all be effective indicators of the concept of intelligence for individual populations: Is she quick-witted?, Does she give considered responses?, Is she good at knowing whom to ask for help?, and Is she good at finding solutions to urgent problems? However, in formulating the most salient questions for each local context and thereby focusing on different kinds of intelligence, the degree of overlap in the construct of intelligence across populations might be greatly reduced (cf.

Brislin, 1986). A further drawback is that ADQ designs do not permit the item-for-item comparison that underlies full scalar equivalence. As a result, demonstrating equivalence across populations at pretesting stages and in analysis is more complicated, in particular if multiple countries are involved.

The notions of *emic* and *etic* concepts and emic and etic indicators (questions) are basic to much of the discussion of ADQ models. We note that the terms emic and etic are used differently in different fields (cf. Headland, Pike & Harris, 1990; Serpell, 1990). Simply put, emic questions are population-specific in relevance and etic questions are universal in relevance. In similar fashion, emic concepts are concepts considered salient for one population and etic concepts are considered to be universal. If an ADQ study uses emic questions to tap a construct/concept assumed to be etic and analysis demonstrates this is the case, the literature speaks of a "*derived* etic". When researchers decide to ask the *same* question of different populations, they assume the question has etic status. Here the literature speaks of an imposed etic, reflecting the top-down approach taken. Prominent early advocates of emic-etic approaches were the psychologist Triandis (1972) and the political scientists Przeworski and Teune (cf. 1966, 1970). Brislin (1980) provides a useful discussion of the advantages and potential drawbacks to early emic-etic approaches. Johnson (1998) refers to a number of studies using variations of the emic-etic approach; van Deth (1998) advocates a functionally equivalent approach in deciding which questions to analyze. A recent two-language application is described in Potaka and Cochrane (2004).

Sometimes a mixed emic-etic approach is used, in which a common core of etic questions, shared across countries, is combined with country-specific emic questions to provide better country-specific coverage of the concepts of interest (see, for example, Berry, 1969; van de Vijver, 2003). Finally, we note that ADQ formats are involved in collecting socio-demographic information whenever population-specific formulations are the best option. Sometimes such questions are blends of translation and country-specific formulations. Educational questions asking for highest qualifications, for example, might begin with the same question text (translated) and continue with a list of the qualifications or school types pertinent for a given educational system.

4.6.4 Ask the Same Question

One general drawback in trying to develop shared questions for multiple populations is that the questions may become less specific than would questions designed for a national study. This may result in inadequate coverage of the construct to be measured and in construct bias (cf. van de Vijver, 2003). Country-specific questions can sometimes be added to counteract this, as mentioned earlier in connection with the emic-etic mixed approach. Ask the same question (ASQ) approaches can differ in the degree of cultural input targeted during development. In terms of QoL literature, they might then be described as simultaneous, parallel, or sequential models.

Sequential ASQ approaches: In a sequential ASQ approach, a source

questionnaire is developed and finalized before other versions are produced as translations of the source questions. In this approach, multicultural considerations are basically addressed at the translation stage. The success or failure of an ASQ sequential approach is largely determined by the suitability of the source questions for all the cultures for which versions will be produced. Without cross-cultural input, however, the questions chosen may be culturally biased. Not surprisingly, criticism of sequential ASQ models focuses on the lack of cross-cultural input at the initial stages of development (for example, Skevington 2002; Camfield, 2003; Ponce, Lavarreda, Yen, Brown, DiSagra, & Satter, 2004). Despite such criticism, sequential ASQ procedures are those most frequently adopted in multinational surveys. Questionnaires developed for one context that are translated at some later date for fielding with a population requiring a different language do not count as *designed* for comparative use; they are simply used in different contexts and languages.

Simultaneous ASQ approaches: In a simultaneous ASQ approach, the questions in different languages are generated together. Classic *decentring* is a procedure that produces questionnaires in two languages more or less at the same time. The goal is to arrive ultimately at items in two languages that are felt to correspond without allowing any one language or culture to dominate. Decentring as a question design procedure is not used widely in survey research. However, the term is also sometimes used to refer to adaptation procedures such as discussed in 4.8. Decentring is one of several design procedures that involves the use of translation. Uses of translation to *develop* questions are distinct from translations made simply to produce new language versions needed. These last are discussed in 4.7.

Decentring can begin with existing questions or, alternatively, with a list of concepts for which questions are to be devised. If questions are the starting point, these will change in the process of decentring. As a result, questions cannot be replicated and simultaneously decentred. There are various ways to proceed within classic decentring; we describe only one option here. The procedure for each question can begin in either language:

- A question is devised or chosen in language A and translated into language B. This translation is only the first step towards removing cultural anchoring; thus no emphasis is placed on close translation;
- Multiple paraphrases or further translations are generated for the translated item in language B;
- Paraphrases for the first item are also produced in language A;
- Anything that causes problems in either language with regard to matching or producing a paraphrase or translation is altered or removed. In this way, culturally anchored obstacles are eliminated from the sets of items generated;
- The sets in each language are appraised and the two items considered to match best are chosen as the comparable questions.

Decentring provides researchers with a means of avoiding language and cultural dominance. However, because it removes culturally specific material, it may result in a loss of specificity and saliency. As a result, questions may be less appropriate for fielding in both contexts than emic items would be. As may be

apparent, classic decentring is not suitable for simultaneous production of multiple translations. Apart from the practical difficulty of attempting this process across twelve languages and cultures, construct coverage, indicator saliency and comparability would be at risk.

Parallel ASQ approaches: Parallel models incorporate cross-cultural input in formulating and selecting draft questions. This input can take the form of advance consultation with local experts, their involvement in the drafting group, or strategies such as incorporating questions from all participating countries in the item pool from which source questions are selected. In other respects, the parallel ASQ approach may resemble the sequential ASQ; a source questionnaire is finalized and any other versions needed are produced on the basis of translation.

Parallel ASQ approaches that ensure adequate cross-cultural co-operation at the conceptualization, drafting, and testing stages may offer a viable compromise between the lack of cultural input in sequential approaches and the complex and expensive demands of simultaneous approaches. At the same time, if discussion and testing of the material and questions is conducted in only one language, problems for cross-cultural implementation may be overlooked. Harkness and Schoua-Glusberg (1998), Braun and Harkness (2005) and Harkness, Schoebi, Joye, Mohler, Faass, & Behr (2007) discuss using advance translation as a means to counteract source questionnaire language dominance.

4.7 TRANSLATING SURVEY INSTRUMENTS

Translation plays a key role in most cross-lingual survey projects. Poor translations can rob researchers of the chance to ask the questions they intend and need to ask. At the same time, projects are often reluctant to invest effort, time or funds in translation procedures. This reluctance is sometimes encouraged by bad past experience with professional translators who proved unable to produce the kind of translations needed. Moreover, because survey questions often look deceptively simple, the temptation to do-it-yourself may also be high. A strategy sometimes adopted does without a written translation and instead has bilingual interviewers translate orally whenever necessary.

The important thing to note is that the effort and cost of producing and testing translations are small, compared to the financial investment made in developing and fielding instruments. In contrast, the price to be paid for poor translations can be high. If poorly translated or adapted questions must be discarded at the analysis stage for even one country, these are lost for analysis across all countries.

4.7.1 Current Good Practice for Translation

In the last decade or so, conceptions of best and good practice regarding survey translation have changed noticeably, as have preferred strategies and the technology used. Translation guidelines published by the US Bureau of Census

(Pan & de la Puente, 2005; de la Puente, Pan, & Rose, 2003), by the European Social Survey (Harkness, 2002/2007) and by Eurostat for health surveys (Tafforeau, López Cobo, Tolonen, Scheid-Nave, & Tinto,, 2005) reflect considerable agreement on how to produce and test translated questions. We summarize here key points on which there is growing consensus:

- A range of expertise is needed to produce a successful survey translation product. This includes expertise in survey questionnaire design, substantive understanding of the subject, source and target language competence, translation training and expertise, and knowledge of the local fielding situation. Translators cannot provide all of these;
- Team approaches, such as described below, have been increasingly advocated as a practical way to bring together the necessary competence;
- Translation teams should consist of those who translate, those who review translations and those who take the final decisions on versions (adjudicators). Consultants for specific aspects can be brought in as required (e.g., on adaptation issues).
- Translators should be skilled practitioners who have received training on translating questionnaires and should normally translate out of the source language into their strongest language. Reviewers need to have at least as good translation skills as the translators but should be familiar with questionnaire design principles, as well as the study design and topic. Adjudicators make final decisions about which translation options to adopt. They do this in cooperation with reviewer and translators, or at least in discussion with a reviewer. Adjudicators must (a) understand the research subject, (b) know about the survey design, and (c) be proficient in the languages involved;
- It is better to use several translators rather than just one, not only in projects where regional variation is expected within the translated language. (cf. Harkness, 2002/2007);
- Wherever feasible, each translator should make a draft translation. The alternative is to have each translator do a section. (See Harkness, 2002/2007; Harkness & Schoua-Glusberg, 1998 on such "split" translation techniques.);
- Translators should be part of the review team and not only employed as translators;
- Translation and adaptation go hand-in-hand (see 4.8);
- Translated questionnaires should be assessed using both quantitative and qualitative procedures (cf. suggestions in Harkness et al, 2004; Harkness & Schoua-Glusberg, 1998; Smith, 2004);
- Translated questionnaires should be pre-tested for the intended population;
- Performance and output should be checked at an early stage in the project when feedback can lead to improvement and save time;
- Team members should be briefed on tasks and responsibilities.

For translators this may include briefing on questionnaires and applications, the mode of data collection, the target audience and required level of vocabulary (see Harkness, 2002/2007). Reviewers should be briefed on their role in reconciling the requirements of question design and those of translation as well as on monitoring translation output. Adjudicators may need to be briefed on the potential and the limitations of translation as a procedure. All may need clarification on types of adaptation (see 4.8):

- Translators and reviewers should take notes on any points of deliberation to inform review and adjudication and to facilitate version documentation;
- Documentation tools should be used to facilitate review and adjudication. These tools often combine translations, source text and note-taking in one document. Examples are provided on the web;
- Translation costs and time should be explicitly included in the study design and budget;
- The planning for translation should identify all the components that may require translation.

Apart from instruments themselves, descriptions of research projects, information leaflets, interviewer manuals, technical fielding reports, pretesting schedules, focus group reports or schedules, and responses to open-coded questions may require translation.

4.7.2 How A Team Approach Works: The Example Of TRAPD

Translation procedures in the ESS comprise a five-step iterative process of **T**ranslation, **R**eview, **A**djudication, **P**re-testing and **D**ocumentation (TRAPD)[1]. Much of the work leading to a final translation is a team effort. Those involved take one or more of the three different roles mentioned earlier: translator, reviewer, and adjudicator. Consultants are recruited as necessary. Approaches of this sort often merge review and adjudication wholly or in part, depending on the expertise of the team and on practical and logistical considerations. The main steps and strategies are presented later; a detailed account, also dealing with sharing languages and splitting translations, is available on the ESS website provided earlier.

- ESS countries are usually required to produce two draft translations. Each translator produces a draft translation independently;
- At a review meeting, translators and a reviewer go through the questionnaire question by question, discussing translated versions and agreeing on a review version;
- Translators and reviewers take notes on unresolved issues and on any compromise decisions;

[1] Pretesting and documentation steps of TRAPD are not fully implemented in the ESS. Participating countries do not pretest the draft source questionnaire, only their translated versions of the finalized source questionnaire. The opportunity to change source questions is thus restricted. The degree of documentation provided by countries on translation also varies in the ESS.

- Adjudication can be part of the review process, in which case the adjudicator attends the review session. Alternatively, adjudication is undertaken at a further meeting between reviewer and adjudicator, possibly with consultants and translators attending;
- Adaptations a country wishes to make in its translation have to be approved by the central co-ordinator of the ESS;
- Countries sharing a language are encouraged to collaborate after they produce their national draft translation(s). In this way, country A can benefit from solutions found by country B. Unnecessary differences can also be avoided.

For more information on team approaches to translation see Harkness and Schoua-Glusberg (1998), Harkness et al (2004) and, explicitly on the ESS, Harkness (2002/2007).

4.7.3 Back Translation

The homepage of the Australian Institute of Interpreters and Translators (AUSIT[1], 2007), the Australian national association for the translating and interpreting professions, has this to say about back translation: "Contrary to popular opinion, having someone translate a translation back into its original tells you nothing about the quality of the first translation. There are better ways to find out whether you're getting what you paid for." The history of back translation and how it came to be the most frequently mentioned survey translation assessment procedure is complex. As described in 4.6.4, decentring uses a form of back and forth translation and paraphrase to develop questions, although not to assess translations as such. This may explain why back translation is often but incorrectly referred to as a translation approach. Whatever the reason, in the social and behavioral sciences back translation is used primarily as a procedure to assess translations.

At its very simplest, the idea is that by translating the target translation back into the source language researchers can compare two versions in a language they understand (the source language version produced in the back translation and the original source language version) and decide on that basis about the quality of a translation in a language they do not understand.

Currently, back translation is the issue on which guidelines possibly differ most. The ESS only mentions back translation in passing, whereas the US Bureau of Census explicitly states it does not recommend back translation. The International Test Commission is less positive about back translation, as reflected in keynote presentations at the 2006 International Test Commission conference in Brussels. The Eurostat guidelines on health surveys mentioned earlier (Tafforeau et al., 2005) recommend back translation but also note that views on its usefulness differ. Somers (2005) discusses how even back translated machine translations do not indicate whether the quality of the first translation is good or not. One example of commercial company statements on the pitfalls of back translation can be found on the Barinas Translation

[1] http://www.ausit.org/eng/showpage.php3?id=648. Accessed July 2007.

Consultants website.[1]

Early advocates of back translation suggested it was a useful assessment tool but were careful to also mention that it had limitations, even if, in our view, such comments reassert a basic usefulness (e.g., Brislin 1970, 1976 and 1986). Throughout the years researchers have expressed misgivings about back translation (Geisinger, 1994; McKenna and Doward, 2005). Recent criticism has emphasized that, since the target language text is the real object of interest, review procedures should focus on this text and not source language texts (Harkness & Schoua-Glusberg, 1998). At the same time, the frequency with which back translation is mentioned in the literature makes it difficult for researchers not to be seen adhering to what has become received practice. As a result, quite elaborate procedures have developed around back translation; either further detailing back translation procedures or adding other assessment procedures before and after back translation (e.g., Sperber , DeVellis, & Boehlecke, 1994; de Mello Alves, Chor, Faerstein, De Lopes,, & Guilherme, 2004).

Although back translation is not a procedure suited to finding subtle but important differences between questions, only targeted research can properly identify which assessment procedures are most useful in which contexts. Targeted research projects comparing back translation with other strategies will doubtless be needed to clarify the effectiveness and costs of alternatives available.

4.8 ADAPTING SOURCE QUESTIONS IN COMPARATIVE CONTEXTS

In terms of source question design, adaptation is the second most popular strategy after replication. In this instance, existing questions are modified and used as the source questions for translations. Such adapted questions are new questions and need to be treated and tested as such.

While translation always involves some kinds of adaptation, adaptation does not necessarily involve translation. In this section, we discuss adaptations of translated questions, not adaptations to source questions. These adaptations are triggered by the act of switching languages, and not by differences in the sociocultural settings and populations.

Educational testing and health research have paid more attention to certain forms of instrument adaptation than have other disciplines (see, for example, Hambleton, 2005; Cook, Schmitt-Cascallar, & Brown, 2005; Chrostowski & Malak, 2003). In fact, the International Testing Commission Guidelines for test adaptation prefer the term adaptation rather than translation because it is "broader and more reflective of what should happen in practice when preparing a test that is constructed in one language and culture for use in a second language and culture. …Test translation is only one of the steps…and…adaptation is often a more suitable term than translation to

[1] http://www.barinas.com/myths.htm

describe the actual process…" (Hambleton, 2005, p. 4). At the same time, no discipline has developed either a systematic analysis of the kinds of adaptation needed for instruments or a detailed description of the strategies that can be used to adapt and test appropriately. In the following paragraphs we present simple examples of some the basic forms of adaptation encountered in comparative instrument-based research (cf. Harkness, 2004, 2006).

Language-driven adaptation: Because translation entails change, all translated questions are in some sense adapted questions. Thus recommendations to keep things the same in translation are bound to fail. Words change, sentence structure changes, the organization of information changes, sound systems change, alphabets change, and the frequency of occurrence of sounds letters and words changes. Comparative linguistics abounds with discussion of differences and similarities between languages. Strictly language-driven changes are fairly predictable instances of adaptation. For example, the English *twenty-eight* is "eight and twenty" in German. Such lexical and structural differences across languages can pose problems for comparability. Thus achieving a good rendering of a source question that accommodates language-driven change and maintains required measurement properties is often a major challenge.

Sociocultural, system-driven adaptation: Measurement systems are a good example of this kind of adaptation (yards, pounds, fahrenheit vs. metres, kilos, centigrade), as are functionally equivalent institutions (parliamentary elections, primary school, Value Added Tax vs. presidential elections, grade school, and purchase tax). Depending on the purpose of a question, adaptations might be simple or complex. Some, such as distance measurements in inches or centimeters, could be directly calibrated if that were necessary or roughly matched if that were sufficient. Hanh et al. (2005) report that the Adolescent Duke Health Profile question *Can you run 100 metre*s? was adapted for Vietnam to ask *Can you run 100 metres or the distance between 3 light poles?* The Vietnamese researchers were uncertain that respondents would understand the distance correctly and offered a locally salient approximation. Whereas light poles were the adaptation for Vietnam, something different might be required for rural Africa (for further examples, see Harkness, 2004).

Adaptation to maintain or reduce level of difficulty: Educational tests are biased if it is easier for one population to have access to the knowledge tested or perform the task required than it is for another population of equal ability. Knowledge questions are thus sometimes adapted to maintain the same level of difficulty across different populations. Language-based memory and vocabulary tests also need to accommodate differing average lengths of words and the relative frequency and difficulty of words chosen across languages. Depending on the test, other aspects, such as ease of pronunciation or visual complexity, might bias recall, repetition or interpretation. In social science, reducing respondent burden is more the issue; adjustments are thus often made to the level of vocabulary used in a translation for populations with expected low levels of education.

Adaptation to ensure local coverage of a concept: Health research has become increasingly cognizant of the fact that translated questions may not ask for the local information needed to ascertain the presence of a given medical

condition (Rogler, 1999; Cheng, 2001; Bolton, 2001; Andary, Stolk, & Klimidid, 2003). The 2000 version of the Diagnostic and Statistical Manual of Mental Diseases, for example, includes localized indicators for depression not present in earlier versions (Cheung, 2004). Similar needs of local or localized questions to improve construct or concept coverage could be identified for many areas—for political or social commitment, religious identification or environmental perceptions and behaviors.

Adaptation to ensure questions are understood as intended: Vision assessment questions are sometimes formulated along the lines of *Do you have difficulty reading a newspaper, even with spectacles?* Such questions assume that respondents are literate, that is, can read, have access to newspapers and, if their vision is impaired, also have access to corrective aids. Someone who is illiterate, for example, might understand the questions as one about whether they know how to read. If newspapers or access to eye care are not readily accessible, other unintended readings of the question could become salient. The question would thus need to be adapted or possibly reframed entirely.

Adaptation related to cultural norms of communication and disclosure: Speech communities differ in how they frame and conduct communication. Depending on cultural expectations regarding politeness, more or less overt expressions of politeness may be required (polite imperatives, apologies for asking a question, etc). A question about female personal hygiene, for example, begins in Asian countries with an apology for asking the question. This is not found in the corresponding English question. In a similar fashion, populations unfamiliar with the survey question and answer game may need more explanation and more directions about what to do than survey-savvy populations would.

Adapting design features: Changes in the design of an instrument can be motivated by many factors including a number mentioned earlier. The direction languages are read or written in, familiarity with certain visual representations (thermometers, faces) and an array of culturally anchored conventions related to visual presentation including color symbolism may call for design adaptation (cf. Tanzer, 2005 on diagram processing). Lexicon (a language's vocabulary) and grammar may also motivate a change in design. For example, the English mid-scale response category *neither agree nor disagree* is rendered in Hebrew ISSP questionnaires as "in the middle". A word-for-word equivalent of the English *neither agree or disagree* in Hebrew would produce "no agree no no agree". Because this means as little in Hebrew as it does in English, a functionally equivalent label is used instead. As things stand, little is known about the effects of changing response scale formats across languages.

The examples presented here are simple; adaptation issues can quickly become quite complex. If information about adaptations and the rationale behind them were drawn together in a databank, it would be possible to learn more about regularities in adaptation needs. In this way a typology could gradually be developed for different disciplines. A cognitive testing report databank called Q-Bank that is being developed by U.S. government agencies could serve as a model for such work. Longer term, such information on adaptations could inform revision and adaptation practices.

4.9 CONCLUSION AND OUTLOOK

The volume of comparative survey research has been growing for decades and the need for global data has never been greater. It is hard to imagine a field which does not use survey data in one form or another. As comparative research projects become increasingly ambitious, technological developments in applications and documentation have increased the power of tools and reduced the effort involved. At the same time, the methodological research needed to inform essential procedures for comparative research has not yet been systematically addressed. As Harkness and colleagues (2003) note, comparative research challenges described in literature of vintage date have still not been systematically addressed.

This chapter focused on important issues for which answers, partial or not, must still be found. There is a good sense in some sections of the research community about what the key comparative methodological issues are and how these might be tackled. A number of the problems faced are, in fact, problems shared across disciplines. On these fronts an increase in cross-disciplinary exchange and collaboration could markedly accelerate progress. Initiatives on different aspects of comparative research could, for example, pool findings, and benefit mutually.

Research on modes in survey research programs such as the ESS and the ISSP could also be shared, as could the work in the ISSP methods work groups on demographic variables, on translation, and on question design. By testing hypotheses and methodological procedures empirically and by ensuring that knowledge and skills accrued are widely shared, progress can be made on issues discussed for over three decades. Joining forces would help groups to find resources to conduct much-needed methodological research. The guideline initiatives in the International Workshop for Comparative Survey Design and Implementation (www.csdi-workshop.org) and the International Test Commission spring to mind as examples.

Standards and protocols developed in one project can serve as models for others. The funding provided by the European Union and participating countries for the ESS, for example, has made it possible to develop protocols and good practice procedures that can benefit other projects, irrespective of whether they adopt the same tactics. In fact, some of the procedures developed in the ESS can be traced back to experience gained in the ISSP. The EU clearly recognizes the importance of evidence-based methods for comparative research. An infrastructure grant to the ESS, for example, has funded training, research, and dissemination projects. As research becomes available that will change expectations and establish new standards, it is critical that research communities collaborate and share their individually developed techniques and expertise. It is also important to avoid a situation in which deserving but modestly funded projects find their achievements overshadowed by the prominence of better-funded projects.

Awareness of the need for research on and refinements of comparative survey methodologies is uneven across disciplines and geographical areas. Lyberg (2006) indicates that official statistics in Europe, for example, has not yet shown a sustained interest in comparative survey methodology or

cooperation with other fields. Certain parts of the world have only very modest survey infrastructures and limited access to training, literature, or basic tools for their work. Survey research is also not welcomed in every part of the world, although national needs for data on topics such as household composition, migration, education, health, and human capital encourage governments to promote data collection and dissemination.

There are also areas in which cross-national, cross-cultural research very much needs to recognize and incorporate methodological advances made in national centers of excellence. At the same time, research across countries or within countries has its own special requirements and procedures. Comparative research is not simply an elaborate extension of general survey research. Certain core challenges, such as question design, are both complex and in some respects politically charged. Commitments to existing instruments, for example, and the time series these represent make it at times difficult to introduce new questions or new design approaches.

Notwithstanding, recent developments, this suggests that considerable methodological progress is likely in the coming decade. These include the ESS infrastructure projects, the ongoing success of the ISSP program and its methodological activities, the emergence of CSDI and CSDI work groups and the international methods conference and monograph planned by that group for 2008, the growth in thematic sessions on comparative research at conferences, the increase in the number of courses taught on comparative survey research in a variety of places, the establishment of the European Survey Research Association (ESRA) and the appearance in 2006 of *Survey Research Methods*, an online journal focusing on methodological issues. Comparatively speaking, the future is most, encouraging.

GLOSSARY OF KEY CONCEPTS

Adaptation. Adapted questions are derived from existing questions by deliberately changing some content or design component to make a question more suitable for a new sociocultural context or for a particular population. Adaptation can be necessary without translation being involved (e.g., adapting a questionnaire for children). However, whenever translation is necessary, some forms of adaptation are also generally required. Adaptations may be substantive, relate to question design, or consist of slight formulation and wording changes. Regardless of the form or the degree of change, it is wise to consider adapted questions as new questions and to test them accordingly.

Ask-Different-Questions Approach (ADQ). In ADQ approaches, researchers collect data across populations/countries using the most salient population-specific questions on a given topic that are felt or demonstrated to tap a construct that is germane or shared across populations.

Ask-the-Same-Question Approach (ASQ). With the exception of *decentring*, researchers adopting ASQ approaches collect data across populations/countries by first deciding on a common source questionnaire in one language and then producing whatever other language versions are needed on the basis of

translation. Although *close translation* is often preferred, adaptations of several kinds may nonetheless be necessary.

Back Translation. Back translation is a procedure which can be sued for several purposes but in survey research is now most often used to assess translations. The translated questionnaire is translated back into the source questionnaire language. Then these two versions in the source language are compared for difference or similarity. Good similarity between these two is taken to indicate that the translated text, which is not itself examined, is faithful to the original source questionnaire.

Close translation. A variety of terms, including close translation, are sometimes used to express that a translation tries to stay as close as possible to the original text in content, presentation and in the case of surveys, format and design. In practical terms, a close translation policy often stands at odds to an approach embracing *adaptation*.

Decentring. In classical decentring models, two different cultures are asked the same questions but the questions are developed simultaneously in each language. Thus there is no source questionnaire or target language questionnaire. The decentring process removes culture-specific elements from both versions. Decentring can thus be seen to stand between *ADQ* models and models based on *ASQ* source questionnaire and translation models.

Etic-Emic. Following distinctions developed by Pike, etic concepts or constructs are universal and therefore shared across multiple cultures, whereas emic concepts or constructs are culture-specific in constellation or significance and cannot be assumed to be shared across populations.

Functional Equivalence. Multiple definitions of functional equivalence exist within and across disciplines. When used in this chapter, it refers to the comparability of the function of a question in a specific context with that of another question in a different specific context.

Team translation. A team translation approach as used in this chapter, combines translation with translation review. It (a) uses more than one translator (b) involves the translators in the review process and not just for the first stage of draft translation (c) brings other expertise to the review process (e. g., survey design and implementation, substantive) and (d) reiterates translation, review, adjudication, and testing as necessary. Thus a good part of the work is carried out by members of the team working as a group.