Journal of Official Statistics, Vol. 19, No. 2, 2003, pp. 153–176

Prevention and Treatment of Item Nonresponse

Edith D. de Leeuw[1], *Joop Hox*[2], *and Mark Huisman*[3]

Item nonresponse is a problem that frequently occurs in survey data. Although it can never be prevented totally, it can be reduced considerably. This not only provides researchers with more data to use in their analyses, but also with helpful auxiliary information for better imputation and adjustment. Understanding, prevention, and imputation are chained together in handling missing data successfully. To supply survey researchers with adequate tools we will review all three aspects. First, we present a typology of missing data patterns and their origins. Based on this typology and the potential sources of item nonresponse, we outline how missing data can be prevented as much as possible. Finally, we discuss how knowledge of the data collection process can improve the statistical treatment of the remaining missing data.

Key words: Causes of missingness; data collection mode; ignorability; imputation; item nonresponse; questionnaire development; follow-up survey.

1. Introduction

An important indicator of data quality in surveys is the amount of item missing data (Groves 1989). When item nonresponse occurs, a unit (e.g., a person) provides data, but for some reason data on particular items or questions are not available for analysis (Dillman, Eltinge, Groves, and Little 2002). Not so long ago, researchers solved this problem by ignoring it and restricting the analysis to observed values or to complete cases. However, this results in loss of information; estimates will be less efficient, and statistical tests will have less power. Furthermore, there is the possibility of systematic differences between units that respond to a particular item and those that do not respond (Rubin 1976). In that case, the basic assumptions necessary for pairwise or listwise deletion are not met and the results of the analyses may be severely biased. A practical problem is that analyses may be performed on different data (sub)sets, and therefore can be inconsistent with each other (Lessler and Kalsbeek 1992; Huisman 1999).

Modern strategies to cope with missing data are imputation (Little and Rubin 1987; Rubin 1987) and direct estimation (Arbuckle 1996; Vermunt 1996). Imputation replaces

[1] MethodikA, Plantage Doklaan 40, NL-1018 CN Amsterdam, The Netherlands. Email: edithl@xs4all.nl
[2] Utrecht University, Department of Methodology and Statistics, P.O.Box 80 140, NL-3508 TC Utrecht, The Netherlands. Email: j.hox@fss.uu.nl
[3] University of Groningen, ICS/Statistics & Measurement Theory, Grote Kruisstraat 2/1, NL-9712 TS Groningen, The Netherlands. Email: m.huisman@ppsw.rug.nl
Acknowledgment: The authors thank the referees of JOS and the Associate Editor David Marker for their stimulating and constructive comments.

the missing values with plausible estimates, to make the data set complete. Direct estimation means that all available (incomplete) data are analyzed using a maximum likelihood approach (cf. Section 5.2). The increasing availability of user-friendly software will undoubtedly increase the use of both imputation and direct estimation techniques. However, a prerequisite for the statistical treatment of missing data is that more is known on how and why the missing data occur. For instance, a missing value that originates from the purely accidental skipping of a question differs from a missing value that originates from the reluctance of a respondent to reveal sensitive information. The first type of missingness can be seen as missing completely at random, the second type of missingness is not random and requires more sophisticated techniques that include a model for the missingness. Thus, researchers find themselves in a "catch-22" position: one should know more about the missing data to select the best method for treating the missing data. Therefore, prevention is a necessary first step in handling missing data. Reduction of item nonresponse will lead to less imputation in a data set, to more data to investigate patterns of item nonresponse and select the best treatment, and finally to more data to base a correct imputation on. Survey methodologists and fieldwork experts (the reducers) can supply survey statisticians and imputation experts (the adjusters) with useful information to base their adjustments on, provided they have taken precautions and planned for nonresponse (Groves and Couper 1998). Handling item nonresponse successfully requires the pooled effort and collaboration of "reducers" and "adjusters!"

In this article, missing data in surveys is discussed. We aim at bringing together the knowledge from survey methodologists who focus on nonresponse reduction, and survey statisticians who concentrate on statistical adjustment. The emphasis is on a better understanding of item nonresponse and its treatment. In Section 2, we start by defining several forms of missing data. Based on this typology, determinants of missing data are distinguished in Section 3 and the implications for statistical treatment are discussed. Also, knowing the determinants, actions to prevent nonresponse can be taken, and this is addressed in Section 4. In Section 5 we give an overview of relatively simple adjustment techniques and available software. We end with recommendations for better handling of missing data in surveys.

2. A Typology of Missing Data

Whenever data are collected with questionnaires or interviews, missing data will occur. There are several types of missing data patterns, and each of them can by caused in different ways and by numerous factors. To successfully prevent and treat missing data it is necessary to define and understand the various types of missingness that may emerge, and to discern the actual missing data patterns. The first concern is the randomness or nonrandomness of the missing data (Rubin 1976; Little and Rubin 1987).

2.1. *Missing at random or not missing at random*

A basic distinction concerning missing data is that data are (1) missing *completely at random*, (2) missing *at random*, or (3) *not* missing at random. This distinction is important because it refers to quite different processes, requiring specific strategies in subsequent data analysis (Little and Rubin 1987).

Data are called *missing completely at random* (MCAR) if the missingness of a response to a question is unrelated to its unknown value and also unrelated to the values of responses to other questions – for example when an interviewer during an interview, or a respondent in a mail questionnaire, accidentally overlooks a question. In the case of missing completely at random, the missing values are a random sample of all values and not related to any observed or unobserved variable. Thus, results of data analyses will not be biased, because there are no systematic differences between respondents and nonrespondents, and problems that arise are mainly a matter of reduced statistical power. It should be noted that the standard solutions in many statistical packages, those of listwise and pairwise deletion both assume that the data are missing completely at random (MCAR); a very strong and often unrealistic assumption.

When the missingness is related to the observed data but not to the (unknown) value of the missing response to the question itself, it is said that the data are *missing at random* (MAR). For example, an elderly respondent has difficulty remembering an event because of a deficient memory. The resulting missingness is related to age, not to the event itself. When the data are missing at random the missingness is a random process conditional on the observed data. In other words, the missing values are a random sample of all values within classes defined by observed values. In the example of the elderly respondent, the data are a random sample within subgroups formed by age. If the data are missing at random (MAR) *and* if the proper statistical model is used, the missingness is said to be *ignorable* with respect to a particular type of inference (e.g., likelihood-based or Bayesian; Rubin 1976). For example, in the case of the elderly respondent, the variable related to the missingness (age) is measured and available for inclusion in the proper analysis (e.g., imputation or weighting adjustment).

Finally, when the missingness is related to the unknown (missing) answer to the question itself, the data are *not missing at random* (NMAR). For example, a respondent perceives her real answer as socially undesirable (e.g., drinking a lot) and evades responding by producing a "do not know" or "no answer." If the data are not missing at random, serious bias may occur. In that case, the missingness is said to be nonignorable and no simple solution for treating the missing data exists. A model for the missingness must be postulated and included in the analysis to prevent bias.

To decide which nonresponse mechanism is at work and how to treat the item nonresponse in the statistical analyses, it is important to study the patterns of missing data.

2.2. *Nonresponse definitions and missing data patterns*

The first type of missingness that is distinguished is *unit nonresponse*, when data for a whole unit of analysis are not available for statistical analysis. Complete records may be lost because units could not be contacted or refused to cooperate, or because the questionnaire of a unit that did cooperate got lost during data editing or analysis (Lessler and Kalsbeek 1992). Unit nonresponse falls outside the scope of this article. For a general overview of unit nonresponse, see the special issue of JOS (1999: 2) and the monograph by Dillman, Groves, Eltinge, and Little (2002).

Unit nonresponse is often called first-level nonresponse. *Item nonresponse* is referred to as second-level nonresponse. The unit has participated, but data on particular items are

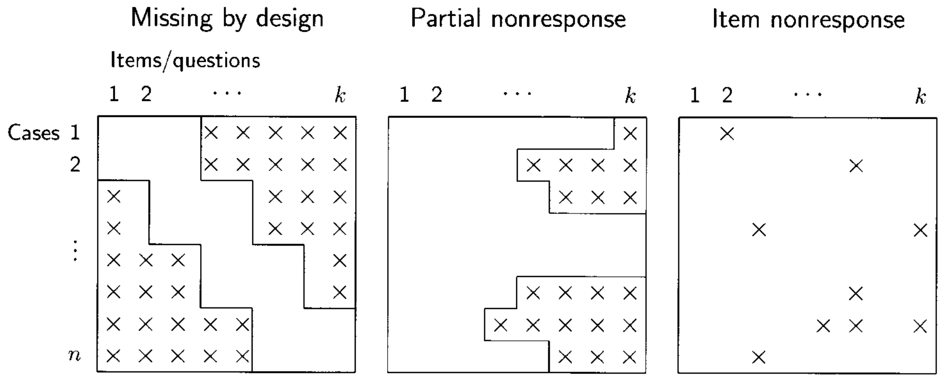

Fig. 1.

unavailable. The term unavailable is used on purpose. Whether or not an answer is counted as missing depends on the goal of the study. For instance, ''do not know'' can be viewed as a meaningful response to a question about voting intentions in an election poll. For other questions (e.g., income), ''do not know'' has no informational value and is counted as missing. Therefore, an item is missing if the researcher interprets it as such, and decides that some kind of treatment (e.g., imputation) is required. Thus, item nonresponse is defined as the failure to obtain *information* for a question in an interview or questionnaire, so *data* are missing (see also Groves 1989).

Three main patterns can be discerned in item missing data: (1) the data are missing systematically by design, (2) all the data are missing after a certain point in the questionnaire (partial nonresponse), and (3) data are missing for some items for some respondents (item nonresponse). These patterns are illustrated in Figure 1.

Each pattern can occur for several reasons, which are discussed in more detail below. For the treatment of missing data, it is important to know how missing data occur. Knowledge of the structure and patterns of the missing data – that is, knowledge of the missing data mechanism – gives us tools to reduce the occurrence of missing data considerably. Moreover, this knowledge is important in selecting the correct treatment procedure. It enables us to analyze the data in an appropriate way by taking the missingness mechanism into account (see Roth 1994, for a discussion of missing data mechanisms).

2.2.1. Missing by design

In the case of missing by design, the researcher decides that certain questions will not be posed to certain persons. There are two main reasons for items to be missing by design.

First, the data are missing because certain questions are not applicable to all respondents, and the questionnaire routing skips these questions. This is also called missing by logic (Huisman and van der Zouwen 1999). For example, one does not pose questions about children to persons without children, or in a health survey one does not ask follow-up questions about a certain disease if the respondent has never had it. The applicability of the questions depends on known characteristics of the respondents, which makes the handling of the missings in the data set relatively easy. Missing by logic occurs because a real answer to the question does not exist and the reason why is known to the researcher.

In other words, values of other questions determine the missingness, thus the missingness mechanism is accessible to the statistical analyst and can be incorporated in the analyses.

The second reason for items to be missing by design is when the missingness is caused by using a specific design to administer different subsets of questions or stimuli to different persons. Contrary to missing by logic, in this case all questions are applicable to all respondents. However, asking all questions would lead to large questionnaires and many hours of completion time. Therefore, specific subsets of questions are posed to different groups of respondents, following well-known designs with known properties. These designs are much used in experimental psychology and have their roots in early agricultural experiments. They have many practical advantages (e.g., less time or resources are needed, the respondent's task takes less time and is less stressful, the data collection is more efficient). A prime illustration is the vignette study on equality of income by Hermkens (1983). Another example of missing by design is the use of incomplete testing designs in educational testing, for instance adaptive or tailored testing (Jansen 1997).

In the case of missing by design, the decision to create missing data by not posing the question is controlled by the researcher. Again, the missingness mechanism is accessible and may be included in the analyses. Thus, using the known properties of the design, the incomplete data can be handled statistically and the analyses give unbiased results. For instance, in the case of vignettes in a factorial survey, multilevel analysis may be used, (for a thorough discussion and example, see Hox, Kreft, and Hermkens 1991). In educational testing some frequently used incomplete testing designs have known missing data mechanisms and can be handled statistically (Mislevey and Wu 1996). Including the known missing data mechanisms in the analysis prevents biased results.

2.2.2. Partial nonresponse

Partial nonresponse is characterized by time dependency. After a certain point in time *all* data are missing. Two well-known examples of partial nonresponse are (1) panel mortality or attrition and (2) break-off during an interview.

In panel surveys part of the initial sample will not respond to subsequent questionnaires or interviews. This is called panel mortality or panel attrition. Researchers are well aware of this phenomenon and methods have been developed to reduce nonresponse and dropout in panels (Freedman, Thornton, and Camburn 1980; Kasprzyk, Duncan, Kalton, and Singh 1989; Laurie, Smith, and Scott 1999) and adjust for it (Van der Pol 1989; Engel and Reinecke 1994). Kalton (1986), Lepkowski (1989), and Rizzo, Kalton, and Brick (1996) discuss ways to treat wave nonresponse in panel surveys. Wothke (2000) and Hox (2000) discuss modeling panel dropout with structural equation models and with multilevel models; both are examples of the direct estimation approach.

Respondents may leave panels for various reasons, but of one thing we may be assured: they do *not* drop out completely at random. If the attrition is not related to the topic of the study (e.g., due to a change of job the respondent has less time), it is a case of missing at random (MAR) and weighting or imputation procedures can be adequately employed. If the attrition is related to the topic, the dropout is not missing at random (NMAR) and a model for the dropout must be included in the analyses to prevent bias. In more technical terms, if the missingness mechanism is accessible, that is, the cause is known and available, and it is included in the analyses, the results will be unbiased (Graham and Donaldson 1993). Therefore, it is of extreme importance to know why people drop out. In panel surveys, usually auxiliary data are available from earlier panel waves. Still, one can only guess at the reasons for leaving the panel. It would be wise to directly ask for these reasons in a special brief exit interview. The data from this exit interview, together with auxiliary information from earlier panel waves, can be used to model the attrition and make the missingness mechanism accessible.

Break-offs mostly occur in telephone interviews. At a certain point in the interview, the respondent stops and disconnects. As a result the remainder of the questions are not answered. When the break-off occurs early in the interview and only a few questions are answered, it is usually treated as a unit nonresponse. When the break-off occurs at the end of the interview and most questions are answered, the remaining unanswered questions are usually treated as item nonresponse. The lack of information in early break-offs makes it difficult to handle this type of nonresponse. When a break-off occurs at the end of the interview, information is available to estimate the missingness mechanism and to take care of it in the analyses.

2.2.3. Item nonresponse

Item nonresponse is characterized by blanks (gaps) in the data set for some persons for some specific questions. But not every blank in the data matrix originates in the same way. We distinguish three forms of item nonresponse. Item nonresponse occurs (1) because the information is not provided by a respondent for a certain question (e.g., a question is overlooked by accident, an answer is not known, a refusal to respond), (2) because the information provided by a respondent for a certain question is not usable (e.g., an answer is not possible, falls outside the range of permissible responses, cannot be coded, or is unreadable), or (3) because usable information is lost (e.g., error in data entry).

There is an important distinction between these three manifestations of item nonresponse. The first two (i.e., information is not provided and information is not usable) originate in the data collection phase, and are often the result of problems in the question-answer process. The third is the result of errors in the data processing phase; information that was available and usable is lost due to errors in data entry, data editing, or coding.

The most problematic form of item nonresponse is the case when a respondent does not provide information, because in this case *different* missing data mechanisms may be at work. In the case of item omission, where the respondent *accidentally* overlooks an item, the data are missing completely at random. In that case, the mechanism is ignorable and almost all simple statistical treatments may be used, even "listwise" deletion. When a respondent is willing but *unable* to respond (Beatty and Herman 2002), the data may be missing at random. For example, if an elderly respondent cannot provide an answer due to faulty memory, the missingness depends on the observed variable age, not on the answer to the question itself, and the item of data is thus missing at random. If the data are missing at random and if the variable related to the missingness is available for analysis, the missingness can be handled adequately with relatively simple solutions (e.g., hot deck imputation, regression imputation). However, in some cases, not being able to respond points to not missing at random; the missingness is related to the (unknown) answer to the question itself. This is the case, for instance, when a respondent does not know the answer in an intelligence test. In the case of *refusal* to respond (e.g., the respondent does not want to give a socially undesirable answer), the missingness is definitely not random and the mechanism is not ignorable. In this case, simple solutions no longer suffice and a special statistical model for the missingness must be postulated.

The second form of item nonresponse, unusable information coded as missing, is generally problematic. The reasons for inadequate scores (e.g., outside the range of possible answers, or nonsubstantive responses) are related to the question format and the real value of the question, pointing to not missing at random. Sometimes the real answer is partly revealed; the missingness mechanism is thus, at least partly, known, and can be incorporated to prevent the analyses from being biased. If this type of missingness can be prevented (e.g., by thorough pretesting), it pays to do so!

Finally, losing information because of errors in coding, editing, or storing is usually not systematic, and normally is MCAR. It arises by accident and is not related to questionnaire and respondent characteristics, so the mechanism is ignorable and the solutions are simple.

Whether item nonresponse will cause bias depends on the way the data are missing, that is, it depends on the missing data mechanism. Knowledge of the causes of the different types of item nonresponse implies knowledge of the missing data mechanisms. Therefore, it is important – both for prevention and for treatment – to assess the causes of missingness and the determinants of item nonresponse.

3. Determinants of Item Nonresponse

When a respondent is confronted with a question and everything goes well, a usable response is given. However, if something goes wrong during the question-answer process, missing data may occur. To distinguish the various determinants of item nonresponse it is important to completely understand what happens during the question-answer process (see Tourangeau 1984; Strack and Martin 1987; Schwarz 1997). First, the respondents have to *understand* the question, that is, they have to determine the intended meaning. If they do not understand the question or the response categories, they may respond with a "do not know" or give no answer. In this case, assuming that the data are missing at random (MAR) is the safest option for handling the missingness.

In the next step, the respondents have to *recall* relevant information from memory: a sometimes difficult cognitive task. For an opinion question, they may either retrieve a previously formed opinion from memory or they may generate an opinion on the spot. For knowledge questions and test items, the relevant knowledge has to be retrieved and combined. In a knowledge test "do not know" is an informative answer and the missing data do not cause any problems. For behavioral questions, the respondent has to recall or reconstruct relevant instances of the behavior from memory and determine whether it occurred during the reference period (e.g., the last three months). If the question asks for usual behavior, the respondent has to decide whether the recalled behavior is representative or whether it reflects a deviation from the usual behavior. So, the next step after retrieval from memory is "computing" a *judgment*. In both phases something may go wrong, resulting in a missing data point. For instance, the respondent does not have the relevant knowledge or has memory difficulties, and this results in a "do not know" answer. The resulting missing data are clearly *not* missing completely at random. They may be missing at random, as is the case when older respondents have more difficulty remembering, or even not missing at random when the memory failure is related to the answer itself: some events may be more easily remembered than others, depending on the saliency of the event, the painfulness, etc.

After a private judgment is formed in the mind of the respondent, she/he has to communicate the answer to the researcher. In the case of a closed question, the respondent has to *format* the judgment to fit the response categories. When open-ended questions are used, the judgment has to be verbalized into a preliminary answer. This again may lead to problems; for instance, the response does not fit the response categories at all. As a result the respondent may refuse to answer, or responds with a "do not know." The missingness is related to the "true" answer to the question and, one may assume, not missing at random.

Before the answer is finally communicated, the respondent may want to *edit* the response, for reasons of social desirability or situational adequacy. Especially with sensitive topics or face-to-face interviews this may be the case. If the real answer does fit the response categories, but is socially undesirable, a respondent may escape a potentially embarrassing situation with a small lie (e.g., I do not know, I have forgotten precisely how) or an outright refusal, which clearly points to not missing at random.

Disturbances in the question-answer process will lead to item nonresponse, and only in rare cases can one safely assume that the data are missing completely at random. In most cases this assumption is not tenable and the risk of bias is considerable. Thus, simple unweighted analyses of the reported data are inappropriate. Groves (1989) distinguishes four potential sources of survey errors: the mode of data collection, the questionnaire itself, the respondent, and the interviewer. Nonresponse is often the result of interaction between two or more sources – for instance the interaction between the questionnaire and the respondent, or the interaction between an interviewer and a respondent. Mathiowetz (1999) gives a clear example of the interaction between questionnaire and respondent: a highly salient question will in many cases improve recall and reduce the "do not know" answers. But even if the question-answer process is followed through without any problem, errors may occur in noting the responses, and in the coding and editing of them by interviewers or researchers during data processing and analysis. This means that answers can be lost (see Section 2.3). Data processors are a fifth potential source of error (Lyberg and Kasprzyk 1997). Each of these five sources of item nonresponse will be discussed in more detail in the next section on prevention.

4. Prevention

4.1. The mode of data collection

There are three major modes of data collection in surveys: the self-administered questionnaire, the face-to-face interview, and the telephone interview. Each mode has advantages and disadvantages regarding data quality and nonresponse. Often methods are mixed within one study –in a panel survey, for example, where the first wave is conducted face-to-face, and subsequent waves either by telephone or through a mail (self-administered) questionnaire. For each data collection mode, computer assisted variants have been successfully developed (De Leeuw and Collins 1997). An extensive meta-analysis by de Leeuw (1992) showed that interview surveys (both face-to-face and telephone) generally result in less item nonresponse than mail surveys. The exception is that self-administered questionnaires give better results when sensitive questions are asked. In general, self-administered questionnaires are experienced as more confidential, and respondents give more honest answers to sensitive questions. Furthermore, in the presence of an interviewer respondents may be reluctant to answer such questions at all. When the topic is sensitive or can be threatening to the respondent, the use of self-administered questionnaires reduces the number of missing data (De Leeuw 1992). However, respondents can make mistakes using a self-administered questionnaire, and inadvertently skip a question. Therefore, the layout of questionnaires is extremely important. For layout rules and examples, see Jenkins and Dillman (1997) and Redline and Dillman (2002).

Using computer-assisted questionnaires prevents such mistakes (Nicholls, Baker, and Martin 1997). In a well-tested computer-assisted interview, all intended questions are asked and routing mistakes are avoided. In addition, range and consistency checks during the interview replace much of the post-editing. This makes it possible to rephrase and pose a question again, and reduces the need to edit an inconsistency into a missing value. This prevents one of the most problematic missing values, those originating in unusable information (cf. 2.2.3).

Comparative studies show that computer-assisted interviewing and computer-assisted self-administered questionnaires result in less item nonresponse than paper and pencil surveys (Nicholls, Baker, and Martin 1997; De Leeuw, Hox and Snijkers 1998; Van Hattum and de Leeuw 1999). It must be emphasized that a ''do not know'' response should not appear explicitly on the screen, instead a special escape key should accept a ''do not know'' if necessary. That computer-assisted self-interviewing (CASI) can be successfully implemented even for special groups, such as children or visually impaired individuals has been demonstrated by De Leeuw, Hox, Kef, and Van Hattum (1997). The development of Audio-CASI makes it possible even for respondents with poor reading skills to report on sensitive topics in a private setting (Turner, Forsyth, O'Reilly, Cooley, Smith, Rogers, and Miller 1998).

4.2. *The questionnaire*

A well-designed questionnaire helps to avoid mistakes of both interviewers and respondents. The importance of questionnaire design for data quality in self-administered questionnaires has been recognized for a long time (cf. Dillman 1978). The same principles govern data quality in interviewer-administered questionnaires. Even good interviewers make mistakes, and routing errors will occur when instructions are not clear. Routing and branching errors can be avoided by a transparent layout, which uses principles of visual perception and graphical design (Jenkins and Dillman, 1997; Redline and Dillman 2002) and guides the interviewer or the respondent error-free from question to question.

Furthermore, the question-answer process must be completed successfully. The question and the question wording must be easy to understand, and the response categories must fit and be exhaustive. In addition to question wording, the number of response categories (the item format) is important. In general, a larger number is better than just two response categories. People often feel uncomfortable with only two forced choices; their intended answer does not fit either alternative and they escape in a "do not know," or cannot answer. The result is data that are *not* missing at random that pose a serious threat of bias. If it is possible to prevent this type of item nonresponse one should do so. Empirical studies have shown that four to seven categories are optimal (Leigh and Martin 1987; Krosnick and Fabrigar 1997), but all response categories should be meaningful to the respondent. For instance, a neutral mid-category is only meaningful when a bi-polar response scale is used (e.g., an agree-disagree response scale). Therefore, a neutral mid-category should only be used in a bi-polar response scale. It should be noted that mode of data collection influences the number of optimal response categories too; in a telephone survey five is a workable maximum (De Leeuw 1992).

An important characteristic of the questionnaire that influences the item nonresponse is the inclusion of "do not know" options (Sudman and Bradburn 1974; Krosnick 2002). There has been much debate about whether or not to include a "do not know" response category (see Schuman and Presser 1981; Gilljam and Granberg 1993). From a practical point of view one should never explicitly offer an easy way out through a "do not know"-category or a no-opinion filter, unless these are substantively important. There is no solid evidence that explicitly offering them is improving the data quality, and they do increase the item nonresponse (Krosnick and Fabrigar 1997; Krosnick 2002). However, a respondent should always be able to answer "do not know" if necessary. For instance, on the interview form or the computer screen a "do not know" option should be available, although the interviewer does not read out this option when reading the response alternatives. The same is true for the option "not applicable." Of course, when "do not know" is a meaningful answer, for instance in reaction to a question on voting intention, it should be offered explicitly. In this case "do not know" has informational value and it does not represent a missing data point!

In a self-administered questionnaire it is wise to omit an explicit "do not know" response category, unless "do not know" is a meaningful answer. In the introduction of the questionnaire, the respondents can be assured that they are allowed to skip a question when they really do not know the answer. In computer-assisted self-interviewing a special function key can be assigned for this purpose (see De Leeuw et al. 1997; Van Hattum and De Leeuw 1999).

Besides layout and question format, clear instructions help to guide the respondent through the question-answer process. Therefore, short and clear instructions to the respondent should be embedded at appropriate places in the questionnaire (Dillman 1978). This helps to prevent missing data. For example, when asking how many times something occurred, "zero" is a meaningful answer. In this case, the instructions should state clearly that one should explicitly fill in "zero" and not skip the question. If this instruction is not given, many respondents may skip the question and a large number of unnecessary missing values be assigned during coding and editing. These missings are related to the true answer and clearly not random (Skinner 1999).

Even the best researcher does not always develop perfect questions. It is therefore essential that questions be tested. Two forms of testing are recommended: the pretest and the pilot or field test. The pretest is an intensive small-scale test in which all steps of the question-answer process are checked. Focus groups or individual intensive cognitive interviews are used to discover problems and to learn how respondents interpret the question. This is usually done after the semi-final version of the questionnaire is completed. A small number of potential respondents is invited to the office and intensive in-depth interviews are conducted, during which the researcher checks whether all steps in the question-answer process are completed successfully or not. Usually, after the pretest, the questionnaire is adapted to facilitate the understanding of key terms. For an introduction into cognitive pretests, see Forsyth and Lessler (1991), Campanelli (1997), and Snijkers (2002); for the use of focus groups in questionnaire development, see Bishoping and Dykema (1999). When preparing for large-scale surveys, a pilot or field-test usually is planned before the main data collection. A pilot is a small-scale realistic test in the field of the total survey, and includes sampling, approaching respondents, data collection, coding, and editing.

4.3. The respondent

Answering questions is a difficult task and respondents may fail to provide adequate responses. Respondents may overlook questions by accident, they may refuse to answer, or they may not be able to provide a correct answer. This can be caused by a problem in the question-answer process (e.g., the respondent does not understand the question or is not able to retrieve the information needed), by lack of motivation on the part of the respondent, by the topic of the question (e.g., sensitive issues), by the structure of the question (open versus closed), by the difficulty of the question, or by badly designed questionnaires (see also, Dillman et al. 2002).

Respondents' characteristics that consistently correlate with item nonresponse are age and education. Elderly respondents and less educated respondents tend to have a higher number of missing data (Colsher and Wallace 1989; Dillman 1978; Herzog and Rodgers 1992; Knäuper, Belli, Hill, and Herzog 1997; Schuman and Presser 1981; Sudman and Bradburn 1974). Because of these consistent correlations the assumption of missing completely at random (MCAR) is not tenable and the standard solution of listwise or pairwise deletion may cause bias. The missing data are at best missing at random (MAR) and at worst not missing at random (NMAR).

There are several ways in which a researcher can attempt to minimize respondent failures. First and most important of all, special attention should be paid to questionnaire, question writing and question testing. This has been extensively discussed above, and only those aspects that are of special importance to the respondent's role in the occurrence of missing data are mentioned here. When testing questionnaires, special attention should be paid to question comprehension and the inclusion of all relevant response categories. A well-tested questionnaire is the basis of good data quality.

Second, mistakes should of course be prevented as far as possible. An ergonomic layout of the questionnaire and instructions embedded in the questionnaire will help the respondent in providing the right answer in the right place (Jenkins and Dillman 1997). A well-written introductory letter and interesting questions can keep a respondent motivated (Dillman 1978), just as a well-trained and attentive interviewer can stimulate a respondent to carefully go through the question-answer process.

Third, computer-assisted self-interviewing methods can help to avoid respondent mistakes or refusals. The interview program takes over the interviewer role and handles the questionnaire logic and questionnaire flow, making it easy for the respondent to answer. The respondent remains the locus of control and determines the pacing of the interview. This gives the respondent more time to understand the question and retrieve and compose an answer, which improves the quality of the answer (Schwarz, Strack, Hippler, and Bishop 1991). Computer-assisted interviewing (CASI and Audio-CASI) also leads to more self-disclosure when sensitive questions are asked (for an overview see Richman, Kiesler, Weisband, and Drasgow 1999).

When respondents are not very willing to part with sensitive information, there are several methods to stimulate them to give a valid answer. These methods are not exclusive and can and should be used in combination. The underlying idea in all these measures is raising the level of respondents' trust. This can be done by making the method as confidential as possible, for instance by combining an interview with a self-administered questionnaire that the respondent seals in an envelope, by using computer-assisted techniques, or by using special techniques like randomized response. For an overview and meta-analysis of randomized response see Lensvelt-Mulders, Hox, and van der Heijden (2002).

When asking sensitive questions, not only the questionnaire, but also the data collection procedure as a whole has to communicate trust and confidentiality to the respondent. For example, one should always give reassurance and explain briefly how information will be handled and what the reason is for asking the questions. However, one should take care not to overdo this or give lengthy reassurances; this can have the opposite effect (Hippler, Schwarz, and Singer 1990). In general, one should avoid the appearance of censurship, formulate questions in a nonjudgmental style, and make response categories as broad as possible (Catanya, Binson, Canchola, Pollack, Hauck, and Coates 1996; Malow, Gustman, Ziskind, McMahon, and St. Lawrence 1998). Especially with sensitive questions there is a high risk that the missingness depends on the true answer of the respondent and is therefore not missing at random. The more one can prevent this, the better.

When using retrospective questions, selective memory plays a role, and respondents may in fact not know the answer (for an overview of the validity of retrospective reports see Schwarz and Sudman 1996). There are several ways in which the respondent can be stimulated to give an adequate response. For instance, one can encourage the respondent to use personal records (e.g., diaries, banking slips). One can also stimulate a more thorough question-answer sequence by using longer introductions or longer questions (Cannel, Miller, and Oksenberg 1981; Scherpenzeel and Saris 1996). The researcher should do everything possible to ensure that the respondent understands the introduction and the question. For instance, several clear short sentences should be used to build a longer introduction, and long sentences and complicated introductions should be avoided.

There are special techniques to prod the respondent's memory and improve recall. These provide the respondent with memory cues, such as calendars on which special dates (e.g., birthdays, anniversaries, holidays) are noted down (Belli 1998; Freedman, Thornton, Camburn, Alwin, and Young-DeMarco 1988). This is known as the time-line follow-back methodology. Other techniques use domain-dependent encoding of memory. By use of extra introductory questions, the respondent is brought back to the situation in which the researcher is interested. The auxiliary questions are followed by the real question, referring to the topic of interest. For example, the respondent is first asked about her/his last car trip (When did that occur? Was it a business trip or for pleasure? Where did you go? Who drove?), and then the central question is asked (Did you stop to buy petrol?) (see Dillman and Tarnai 1991). For an overview of recall error and bias reduction techniques, see Eisenhower, Mathiowetz, and Morganstein (1991).

4.4. *The interviewer*

Interviewers can have a very positive role in reducing nonresponse (Fowler 1991). They can guide the respondent through the questionnaire and explain questions the respondent does not understand. They can also adequately probe a respondent, that is, ask a question again when a respondent is not quite sure of the answer or even not quite sure about giving an answer at all. In standard interview training, interviewers are often instructed to probe once after an initial ''do not know'' or when a respondent hesitates in choosing the best-fitting response category.

However, interviewers may also induce nonresponse. There are several ways in which an interviewer may cause missing data. The interviewer can, for instance, fail to ask the question, or fail to probe a respondent. Interviewers may also fail to record the answer, or may record the answer incorrectly or illegibly. In the latter case, the answer will often be coded as missing during post-interview editing.

There are two causes for the interviewer failures mentioned above. First of all, the interviewer can make a genuine routing error. Second, the interviewer may fail on purpose, because she/he wants to end the interview quickly or does not want to go to too much trouble. By taking a wrong routing on purpose, the interviewer can avoid some long and tedious questioning. Through not probing and just noting the ''do not know'' option the interview will take less time. When missing data result from deliberate routing failures of interviewers, the missingness may be related to the true value of the question and should therefore be regarded as not random. These interviewer tactics are more likely when interviewers are paid by the completed interview, rather than by the hour.

There are several ways in which genuine mistakes can be reduced or even avoided. First of all, interviewers must be trained intensively in the correct procedures (Sanchez 1992; Billiet and Loosveldt 1988). In addition, they should receive specific instructions about the questionnaire at the beginning of each new survey (McCrossan 1991; Carton 1999). Secondly, an ergonomic layout of the questionnaire or interviewer schedule reduces skipping and routing errors (cf. Redline and Dillman 2002) and using computer-assisted interviewing may even avoid these errors completely (De Leeuw 2002). Errors made on purpose are reduced by strict interviewer supervision and monitoring (De Leeuw and De Heer 2002; Fowler 1991), and in computer-assisted interviewing by the use of computer log-files (Couper, Hansen, and Sadovsky 1997).

4.5. *Data processing*

In the design and planning phase of a study, a researcher can take many steps to ensure that the question-answer process proceeds as smoothly as possible, thereby avoiding errors and improving data quality. But even if the question-answer process is followed through without any problem, errors may occur in data entry, coding, and editing the answer during data processing and analysis (Lyberg and Kasprzyk 1997; Federal Committee on Statistical Methodology 2001).

Computer-assisted data collection and integrated systems for survey processing are useful tools in preventing mistakes (Bethlehem 1997). Another tool is quality control during coding and analysis. Total Quality Management is the key to preventing item nonresponse (see also Dippo 1997; Morganstein and Marker 1997).

5. Treatment

5.1. Necessary knowledge of the structure and patterns of missing data

For an optimal treatment of item nonresponse, knowledge of the missing data mechanism is needed to preclude potential nonresponse bias. Knowing how missing data occur may help one to choose the most appropriate treatment procedure. In the first place one should know whether the data are *missing completely at random* (MCAR) or not. Only in the case of MCAR will the analyses not be biased, because there are no systematic differences between respondents who completed the question and respondents who have a missing value for that question.

If the data are not missing completely at random they are either *missing at random* (MAR), or are *not missing at random* (NMAR). If the data are missing at random, it is important to know which variables are related to the missingness. For a proper treatment of the missing data these variables should be included in the adjustment or analysis model. For example, respondents' age and education consistently correlate with item nonresponse, resulting in data that are not missing completely at random. By use of age and education in the adjustment model this is accounted for.

In the worst case when the data are not missing at random, the missingness is related to the unknown (missing) answer to the question. A prime example is refusing to respond (e.g., no answer given because of sensitive nature of the true answer). If the data are not missing at random, there are no simple solutions; a specific model for the missingness must be postulated (cf. Little and Rubin 1987). As ''do not know'' answers are often a polite way to refuse, these pose a serious problem. Only in those cases where it is likely that a respondent really does not know the answer and when the variables causing the ''do not know'' is known and available for analysis, is relatively straightforward treatment possible (e.g., the variable age when an older respondent has difficulty remembering an event).

How to acquire this knowledge? A first step is to inspect the patterns of missingness. This can provide very practical information. For instance, we may find that most of the missing values concern only one variable. If that variable is not central to our analysis, we may decide to delete it, rather than keeping it and creating a problem by having to delete or impute many cases. The same goes for a single case with many missing values. In general, however, missing values are scattered all over the data matrix. In that general case, we would like to know if the missingness forms a pattern, and/or if it is related to some of our observed variables. If we discover a system in the missingness pattern, we may try to include that in our statistical analyses and imputation procedures. Modern software can assist us in this. For instance, SPSS has a specialized module available called

Missing Value Analysis (MVA) that includes a variety of techniques to inspect the pattern of missing data, and the program SOLAS offers the opportunity to inspect the data visually. See Hox (1999) for an overview of software, and Huisman (1999) for a demonstration of several methods to investigate the randomness of missing data in a collection of data sets.

The inspection of missing data patterns, can however, never tell us with certainty whether or not the missingness is independent of the (unknown) value of the variable (question). Extra information is needed to test the *missing at random* (MAR) assumption and determine the causes of item nonresponse. This can be additional information (from other sources than the actual sample), like theory, logic, or prior data (Graham and Donaldson 1993), and reasonable guesses about the mechanism can be made (Little and Rubin 1987). We can gain even more insight in the missing data mechanism by turning to the respondents. In Section 4.2. we described how a pretest and/or a field test will give suggestions about improving the questionnaire and preventing missing values. But a pretest can do more, it can also result in information as to why certain missing values occur. This information will enable the statisticians to choose the appropriate adjustment technique.

If preliminary analyses of the data set point to problematic variables with many missing data, one can do a qualitative ''post''-test on a small number of respondents. This post-test needs specially selected respondents who are similar with respect to background characteristics to the respondents in the survey that produced many or typical item nonresponses. The structure of the post-test would follow the basic guidelines of cognitive interviews, focusing on the issue would answer certain questions and why (not). A second direct way to collect additional information is to follow up respondents with item nonresponse in order to determine the causes of missingness (Graham and Donaldson 1993). This was done by Huisman, Krol, and van Sonderen (1998), who re-approached a subsample of respondents with missing values on items on health scales. In a standard introduction it was explained that some obscurities had emerged while processing the data and the respondent was asked whether some questions could be repeated. When the respondent was unwilling to answer a second time the interviewer was instructed to probe for the reason why. The resulting information was used to determine the causes of missingness; and given the values of the variables related to the missingness, the nonresponse bias was, at least partly, corrected (Huisman, Krol, and van Sonderen 1998).

The direct approach to respondents, either for a qualitative post-test or for an additional re-interview, clearly increases the costs in terms of time and money. A less costly source of additional information is interviewer observations. Interviewers can observe with intelligence and report on the question-answer process and potential disruptions (Loosveldt 1995; Snijkers, de Leeuw, Hoezen, and Kuipers 1999); they can also report on general income level of the neighborhood and other socio-demographic characteristics (Groves and Couper 1998). These observations can provide valuable information as to the causes of missingness, which can also be used in the imputation stage.

5.2. *Simple and effective methods to treat missing data*

One can successfully cope with missing data, but it is not always easy or straightforward. The default options of statistical software are usually listwise or pairwise deletion, or some very simple imputation technique such as mean substitution. These "solutions" are inadequate. Listwise deletion is wasteful because it discards information. Pairwise deletion may result in inconsistent correlation matrices and totals. Both methods are also likely to be biased. Simple imputation may produce biased point estimates and will underestimate true sampling variances, resulting in spuriously small p-values. In addition, one should realize that these techniques are all based on the very strong assumption of missing completely at random (MCAR). As discussed above, this is seldom the case. Therefore, the best policy is first to prevent missing data as much as possible, and then to use an analytical strategy that consists of two steps (Hox 1999): (1) use all available information to investigate the missing data patterns and (2) analyze the incomplete data set performing the necessary (and correct) adjustment for missing data.

When the data can be considered missing completely at random, simple solutions like listwise or pairwise deletion do not result in nonresponse bias. To the extent that the missingness is random, effective methods exist to treat the problem. For instance, analyzing incomplete data conditional on covariate information is a useful procedure to reduce the nonresponse bias. Rubin, Stern, and Vehovar (1995) recommend using the missing at random assumption as a starting point for analysis in large well-conducted surveys.

Two different analytical strategies can be used when the data are assumed missing at random: direct estimation and imputation. Direct estimation means that the incomplete data are fully analyzed using a maximum likelihood approach. Direct estimation requires specialized software, but this is increasingly becoming available. For instance, several programs for structural equation modeling (e.g., Amos, Mplus, Lisrel 8.5) are now able to include incomplete cases in the analysis. Since analysis of (co)variance, multiple regression analysis, and discriminant analysis can all be formulated as a structural equation model, these analyses can now be carried out under the assumption of missing at random. Another example is using multilevel models for the analysis of longitudinal data. In such analyses, the repeated measures are viewed as hierarchically nested within cases (cf. Hox 2002). Since multilevel models do not assume that all measurement occasions are available for analysis, missing data due to panel dropout are not a problem. The incomplete cases are included in the analysis, and provided that the estimation method used is maximum likelihood, the assumption is that data are missing at random (Little 1995).

While direct estimation is powerful, it requires access to and knowledge of specialized software. With (single) imputation, on the other hand, standard software can be used to analyze the data. Imputation fills the gaps in the data set with plausible values. Many imputation methods exist, which differ in the way they define "plausible." The attraction of imputation methods is that they result in a data file that is complete. The analyst can ignore the missingness problem, and proceed to analyze the completed data using any standard method she/he is familiar with.

A problem with imputation is that most simple imputation methods, such as replacing missing values with the overall mean or using regression methods to estimate the missing values, result in biased estimates. However, the popular and reasonably simple *hot deck* method results in unbiased estimates under the assumption of missing at random. In the hot deck method (cf. Brick and Kalton 1996; Marker, Judkins and Winglee 2002), respondents are sorted into a number of imputation classes according to a set of auxiliary variables. Obviously, the choice of auxiliary variables should be based on an analysis of the missingness pattern. Missing values are then replaced by observed values taken at random from the same imputation class.

Two fundamental problems with imputation are: (1) using the information in the observed data to predict the missing values emphasizes the structure in the completed data, and (2) analyzing the completed data set involves using a spuriously large number of cases and thus leads to biased significance tests. Rubin (1987) proposes to solve these problems by using *multiple imputation*: each missing value is replaced by two or more (M) plausible estimates to create M completed data sets. The plausible values must include an error term from an appropriate distribution, which solves the problem of exaggerating the existing structure in the data. Analyzing the M differently completed data sets and combining the estimates into an overall estimate solves the problem of the biased significance test.

In the multiple imputation approach, analyzing M data sets and having to combine the results is cumbersome, but not especially difficult. What is difficult is generating the M data sets in a proper manner. A nonparametric method (available in SAS 8.2 and SOLAS) is to compute for each respondent the propensity to have missing values on a specific variable, group respondents into imputation classes based on this propensity score, and use hot deck imputation with these imputation classes. Parametric imputation methods assume a model for the data, and use Bayesian methods to generate estimates for the missing values. These methods are described in detail by Schafer (1997). Again, it is important that the model for the data generation is very general and includes those variables that are important for predicting either missingness or the variables of interest.

For an overview of several relatively simple imputation techniques we refer to Martin et al. (1996) and Huisman (1999). Multiple imputation, which can properly model the uncertainty in the imputed values, is certainly *not* simple. However, it is made less difficult by the existence of several programs that impute missing values using a choice of different techniques. The freeware program NORM, which Schafer uses generates multiple imputations under the model that the data have a multivariate normal distribution. Schafer (1997) describes similar procedures for other models, including categorical and mixed categorical-normal data, and panel data. These programs, appropriately called CAT, MIX, and PAN, are available for the statistical package Splus; NORM is also available as a stand-alone Windows program. The program SOLAS (Statistical Solutions, Ltd. 1997) can perform propensity score-based imputation. Finally, SAS 8.2 includes two new procedures: PROC MI to generate multiple imputations, and PROC MIANALYZE to analyze and combine the results. For a comparison of available software see Hox (1999) and Horton and Lipsitz (2001). For an accessible discussion of modern imputation methods we refer to Schafer and Olsen (1998).

6. Conclusions and Discussion

With special attention to each phase of the survey, it is possible to reduce the amount of item nonresponse. Well-trained interviewers, appropriate data collection methods, and a well-designed questionnaire, all help to reduce item nonresponse. It is crucial to extensively pretest the questionnaire to detect problems in question wording or in presentation of the questions, which may lead to errors in the question-answer process.

When a large-scale survey is planned, or when an existing survey is redesigned, field tests usually precede the implementation. In a field test or pilot all procedures necessary for a survey are followed through on a smaller scale. This gives researchers an opportunity for a last check on missing data. Statistical analysis of the collected data can provide information on patterns of missingness, which can be useful for a last redesign of the questionnaire. Similarly, interviewer debriefing studies and follow-up interviews with respondents can provide the researcher with valuable information on problems during the data collection and suggest possible solutions.

Knowing how item missing data occur not only helps to improve the data collection procedure, it also helps to choose the most appropriate statistical treatment. For instance, respondents' age and education consistently correlate with item nonresponse, resulting in data that are not missing completely at random. By use of the variables age and education in the statistical adjustment model, this is accounted for.

Prevention and adjustment are two sides of one coin. Item nonresponse cannot be totally prevented, and imputation will be necessary during the initial data analysis. The more information one has, the better one can investigate the missingness mechanism and the better one can adjust. Therefore, the best policy is first to prevent missing data as much as possible, and in addition collect auxiliary information (either using pretest and interviewer information, or via special cognitive post-tests or respondent follow-ups); second, a two-step analytical strategy should be used: (1) use all available information to investigate the missing data patterns and (2) analyze the incomplete data set performing the necessary (and correct) adjustment for missing data.

7. References

Arbuckle, J. (1996). Full Information Estimation in the Presence of Missing Data. In Advanced Structural Equation Modeling. G.A. Marcoulides and R.E. Schumacker (eds). Mahwah, N.Y.: Erlbaum, 243–278.

Beatty, P. and Herrmann, D. (2002). To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse. In Survey Nonresponse. R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 71–85.

Belli, R.F. (1998). The Structure of Autobiographical Memory and the Event History Calendar. Potential Improvements in the Quality of Retrospective Reports in Surveys. Memory, 6, 383–406.

Bethlehem, J.G. (1997). Integrated Control Systems for Survey Processing. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 371–392.

Billiet, J. and Loosveldt, G. (1988). Improvement of the Quality of Responses to Factual Survey Questions by Interviewer Training. Public Opinion Quarterly, 52, 190–211.

Bishoping, K. and Dykema, J. (1999). Towards a Social Psychological Program for Improving Focus Group Methods of Developing Questionnaires. Journal of Official Statistics, 15, 495–516.

Brick, J.M. and Kalton, G. (1996). Handling Missing Data in Survey Research. Statistical Methods in Medical Research, 5, 215–238.

Campanelli, P. (1997). Testing Survey Questions: New Directions in Cognitive Interviewing. Bulletin de Methodologie Sociologique, 55 (special issue on cognitive interviewing), 5–17.

Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981). Research on Interviewing Techniques. In Sociological Methodology. S. Leinhardt (ed.). San Francisco: Jossey-Bass.

Carton, A. (1999). Selectie, Training en Evaluatie van Interviewers binnen een Interviewernetwerk: (An Interviewer Network: Constructing a Procedure to Evaluate Interviewers). Leuven: Garant [In Dutch].

Catanya, J.A., Binson, D., Canchola, J., Pollack, L.M., Hauck, W., and Coates, T.J. (1996). Interviewer and Question Effects on Sex Items. Public Opinion Quarterly, 60, 345–375.

Colsher, P.L. and Wallace, R.B. (1989). Data Quality and Age: Health and Psycho-behavioral Correlates of Item Nonresponse and Inconsistent Responses. Journal of Gerontology, 44, 45–52.

Couper, M.P., Hansen, S.E., and Sadovsky, S.A. (1997). Evaluating Interviewer Use of CAPI Technology. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 267–285.

Dillman, D.A., Eltinge, J.J., Groves, R.M., and Little, R.J.A. (2002). Survey Nonresponse in Design, Data Collection, and Analysis. In Survey Nonresponse. R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 3–26.

Dillman, D.A. and Tarnai, J. (1991). Mode Effects of Cognitive Designed Recall Questions: A Comparison of Answers to Telephone and Mail Surveys. In Measurement Errors in Surveys. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: Wiley, 74–123.

Dillman, D.A. (1978). Mail and Telephone Surveys; The Total Design Method. New York: Wiley.

Dippo, C.S. (1997). Survey Measurement and Process Improvement: Concepts and Integration. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (eds). New York: Wiley, 457–474.

Eisenhower, D., Mathiowetz, N.A., and Morganstein, D. (1991). Recall Error: Sources and Bias Reduction Techniques. In Measurement Errors in Surveys. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, S. Sudman (eds). New York: Wiley, 127–144.

Engel, U. and Reinecke, J. (1994). Panelanalyse: Grundlagen, Techniken, Beispiele. Berlin: Walter de Gruyter [In German].

Federal Committee on Statistical Methodology (2001). Measuring and Reporting Sources of Errors in Surveys. Statistical Policy Working Paper 31. Washington DC: Executive Office of the President of the United States, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Forsyth, B.H. and Lessler, J.T. (1991). Cognitive Laboratory Methods: A Taxonomy. In Measurement Errors in Surveys. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: Wiley, 393–418.

Fowler Jr., F.J. (1991). Reducing Interviewer Related Error Through Interviewer Training, Supervision and Other Means. In Measurement Errors in Surveys. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: Wiley, 259–278.

Freedman, D.S., Thornton, A., and Camburn, D. (1980). Maintaining Response Rates in Longitudinal Studies. Sociological Methods and Research, 9, 87–98.

Freedman, D.S., Thornton, A., Camburn, D., Alwin, D., and Young-DeMarco, L. (1988). The Life History Calendar: A Technique for Collecting Retrospective Data. In Sociological Methodology. C. Clogg (ed.). San Francisco: Jossey-Bass, 37–68.

Gilljam, M. and Granberg, D. (1993). Should We Take Don't Know for an Answer? Public Opinion Quarterly, 57, 348–357.

Graham, J.W. and Donaldson, S.I. (1993). Evaluating Interventions With Differential Attrition: The Importance of Nonresponse Mechanisms and Use of Follow-up Data. Journal of Applied Psychology, 53, 383–392.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: Wiley.

Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (2002). Survey Nonresponse. New York, Wiley.

Groves, R.M. and Couper, M.P. (1998). Nonresponse in Household Interview Surveys. New York: Wiley.

Hattum, M. van and de Leeuw, E.D. (1999). A Disk-by-mail Survey of Teachers and Pupils in Dutch Primary Schools: Logistics and Data Quality. Journal of Official Statistics, 3, 413–429.

Hermkens, P.L.J. (1983). Oordelen over de rechtvaardigheid van inkomens. (Judgements on the Fairness of Income). Amsterdam: Kobra [In Dutch].

Herzog, A.R. and Rodgers, W.L. (1992). The Use of Survey Methods in Research on Older Americans. In The Epidemiological Study of the Elderly. R.B. Wallace and R.F. Woolson (eds). Oxford: Oxford University Press.

Hippler, H.-J., Schwarz, N, and Singer, E. (1990). Der influess von Datenschutzzusagen auf die teilnamebereitschaft an Umfragen. (The Influence of Data Protection Reassurance on the Willingness to Participate in a Survey.) ZUMA-nachrichten, 27, 54–67. [In German].

Horton, N.J. and Lipsitz, S.R. (2001). Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. The American Statistician, 55, 244–254.

Hox, J.J. (1999). A Review of Current Software for Handling Missing Data. Kwantitieve Methoden, 62, 123–138. (Available online at http://www.fss.uu.nl/ms/jh).

Hox, J.J. (2000). Multilevel Analysis of Grouped and Longitudinal Data. In Modeling Longitudinal and Multilevel Data. T.D. Little, K.U. Schnabel, and J. Baumert (eds). Mahwah, NJ: Lawrence Erlbaum Associates, 15–32.

Hox, J.J. (2002). Multilevel Analysis: Techniques and Applications. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J.J., Kreft, I.G.G., and Hermkens, P.L.J. (1991). The Analysis of Factorial Surveys. Sociological Methods and Research, 19, 493–510.

Huisman, M. (1999). Item Nonresponse: Occurrence, Causes, and Imputation of Missing Answers to Test Items. M and T Series 32. Leiden: DSWO-press.

Huisman, M., Krol, B., and Sonderen, F.L.P. van (1998). Handling Missing Data by Re-approaching Nonrespondents. Quality and Quantity, 32, 77–91.

Huisman, M. and Zouwen, J. van der. (1999). Item Nonresponse in Scale Data from Surveys: Types, Determinants, and Measures. In Item Nonresponse: Occurrence, Causes, and Imputation of Missing Answers to Test Items. Leiden: DSWO-press, 63–90.

Jansen, M.G.H. (1997). The Rasch Model for Speed Tests and Some Extensions with Applications to Incomplete Designs. Journal of Educational and Behavioral Statistics, 22, 125–140.

Jenkins, C.R. and Dillman, D.A. (1997). Towards a Theory of Self-administered Questionnaire Design. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (eds). New York: Wiley, 165–196.

Journal of Official Statistics (1999). Special Issue on Survey Nonresponse, 127–350.

Kasprzyk, D., Duncan, G.J., Kalton, G., and Singh, M.P. (1989). Panel Surveys. New York: Wiley.

Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys. Journal of Official Statistics, 2, 303–314.

Knäuper,B., Belli, R.F., Hill, D.H., and Herzog, A.R. (1997). Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality. Journal of Official Statistics, 13, 181–199.

Krosnick, J.A. (2002). The Causes of No Opinion Responses to Attitude Measures in Surveys: They Rarely Are What They Appear to Be. In Survey Nonresponse. R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 87–100.

Krosnick, J.A. and Fabrigar, L.R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (eds). New York: Wiley, 141–164.

Laurie, H., Smith, R., and Scott, L. (1999). Strategies for Reducing Nonresponse in a Longitudinal Panel Survey. Journal of Official Statistics, 15, 269–282.

Leeuw, E.D. de. (1992). Data Quality in Mail, Telephone, and Face to Face Surveys. Amsterdam: TT-publikaties.

Leeuw, E.D. de. (2002). The Effect of Computer-Assisted Interviewing on Data Quality: A Review of the Evidence (CD-Rom). In Social Science Methodology in the New Millennium. J. Blasius, J. Hox, E. de Leeuw and P. Schmidt (eds). Opladen, FRG: Leske + Budrich.

Leeuw, E.D., de and Collins, M. (1997). Data Collection Method and Survey Quality: An Overview. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: Wiley, 199–220.

Leeuw, E.D., de and Heer, W. de. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In Survey Nonresponse. R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 41–54.

Leeuw, E.D. de, Hox, J.J., and Snijkers, G. (1998). The Effect of Computer-assisted

Interviewing on Data Quality. In Market Research and Information Technology. Application and Innovation. B. Blyth (ed.). Esomar Monograph 6, Amsterdam: Esomar, 173–198.

Leeuw, E.D. de, Hox, J.J., Kef, S., and Hattum, M. van (1997). Overcoming the Problems of Special Interviews on Sensitive Topics: Computer Assisted Self-interviewing Tailored for Young Children and Adolescents. Sawtooth Software Conference Proceedings. Sequim: WA: Sawtooth.

Leigh, J.H. and Martin, C.R. (1987). Do-not-know Item Nonresponse in Telephone Surveys: Effects of Question Form and Respondent Characteristics. Journal of Marketing Research, 24, 418–424.

Lensvelt-Mulders, G., Hox, J., and Heijden, P. van der. (2002). Meta-analysis of Randomized Response Research: 35 Years of Validation. Research Paper, Utrecht University.

Lepkowski, J.M. (1989). Treatment of Wave Nonresponse in Panel Surveys. In Panel Surveys. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh (eds). New York: Wiley, 348–374.

Lessler, J.T. and Kalsbeek, W.D. (1992). Nonsampling Error in Surveys. New York: Wiley.

Little, R.J.A. (1995). Modeling the Drop-out Mechanism in Repeated Measures Studies. Journal of the American Statistical Association, 90, 1112–1121.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Loosveldt, G. (1995). The Profile of the Difficult-to-interview Respondent. Bulletin de Methodologie Sociologique, 48, 68–81.

Lyberg, L. and Kasprzyk, D. (1997). Some Aspects of Post-survey Processing. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 353–370.

Malow, R.M., Gustman, S.L., Ziskind, D., McMahon, R, and St. Lawrence, J. (1998). Evaluating HIV Prevention Interventions Among Drug Abusers: Validity Issues. Journal of HIV/AIDS Prevention for Adolescents and Children, 2, 21–40.

Marker, D.A., Judkins, D.R. and Winglee, M. (2002). Large-scale Imputation for Complex Surveys. In Survey Nonresponse. R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 329–342.

Martin, J., Astill, S., Boothroyd, R., Elliot, D., Hickman, S., Johnson, S., Laux, R., Loyd, R., Macafee, K., Morgan, Ch., Semmence, J., Smith, P., Teague, A., and Skinner, C. (1996). Task Force on Imputation. Report on Imputation, Government Statistical Services, Methodology Series, No. 3. U.K.: GSS/ ONS.

Mathiowetz, N. (1999). Presentation in Response to T. Murato and P.A. Gwartney. Question Saliency, Question Difficulty, and Item Nonresponse in Survey Research. International Conference on Survey Nonresponse, Portland, October.

McCrossan, L. (1991). A Handbook for Interviewers. London: HMSO.

Mislevy, R.J. and Wu, P.-K. (1996). Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing. Research Report RR-96–30-ONR. Princeton, NJ: Educational Testing Service.

Morganstein, D. and Marker, D.A. (1997). Continuous Quality Improvement in Statistical

Agencies. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 475–500.

Nicholls II, W.L., Baker, R.P., and Martin, J. (1997). The Effect of New Data Collection Technologies on Survey Data Quality. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 221–248.

Pol, F.J.R. van de. (1989). Issues of Design and Analysis of Panels. Amsterdam: Sociometric Research Foundation.

Redline, C.D. and Dillman, D.A. (2002). The Influence of Alternative Visual Designs on Respondents' Performance with Branching Instructions in Self-administered Questionnaires. In Survey Nonresponse. R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 179–193.

Richman, W., Kiesler, S., Weisband, S., and Drasgow, F. (1999). A Meta-analytic Study of Social Desirability Distortion in Computer-administered Questionnaires, Traditional Questionnaires, and Interviews. Journal of Applied Psychology, 84, 754–775.

Rizzo, L., Kalton, G., and Brick, M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse. Survey Methodology, 22, 43–53.

Roth, P.L. (1994). Missing Data: A Conceptual Review for Applied Psychologists. Personnel Psychology, 47, 537–560.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–592.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D.B., Stern, H.S., and Vehovar V. (1995). Handling ''Don't Know'' Survey Responses: The Case of the Slovenian Plebiscite. Journal of the American Statistical Association, 91, 822–828.

Sanchez, M.E. (1992). Effects of Questionnaire Design on the Quality of Survey Data. Public Opinion Quarterly, 56, 206–217.

Schafer, J. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Schafer, J. and Olsen, M.K. (1998). Multiple Imputation for Multivariate Missing-data Problems: A Data Analyst's Perspective. Multivariate Behavior Research, 33, 545–571.

Scherpenzeel, A. and Saris, W. (1997). The Validity and Reliability of Survey Questions: A Meta-analysis of MTMM Studies. Sociological Methods and Research, 25, 341–383.

Schuman, H.H. and Presser, S. (1981). Questions and Answers in Attitude Surveys. New York: Academic Press.

Schwarz, N. (1997). Questionnaire Design: The Rocky Road from Concepts to Answers. In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (eds). New York: Wiley, 29–45.

Schwarz, N., Strack, F., Hippler, H.-J., and Bishop, G. (1991). The Impact on Administration Mode on Response Effects in Survey Measurement. Applied Cognitive Psychology, 5, 193–212.

Schwarz, N. and Sudman, S. (eds) (1996). Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer Verlag.

Skinner, C. (1999). Developing an Imputation Strategy, with Illustrations from a Self-completion Survey of Local Authorities. Lecture presented at the Survey Methods

Center Seminar on Item Nonresponse in Surveys. London, Royal Statistical Society, March.
Snijkers, G. (2002). Cognitive Laboratory Methods: Current Best Practices. In Cognitive Laboratory Experiences: On Pretesting, Computerized Questionnaires and Data Quality. G. Snijkers, Utrecht, Ph.D. thesis. University of Utrecht, pp 63–96.
Snijkers, G., Leeuw, E.D. de., Hoezen, D., and Kuipers, I. (1999). Computer-assisted Qualitative Interviewing: Testing and Quality Assessment of CAPI and CATI Questionnaires in the Field. In Leading Survey and Statistical Computing into the New Millennium. R. Banks et al. (eds). Proceedings of the third ASC international conference. University of Edinburgh.
Statistical Solutions, Ltd. (1997). SOLAS for Missing Data Analysis. Cork, Ireland: Statistical Solutions Ltd.
Strack, F. and Martin, L. (1987). Thinking Judging and Communicating: A Process Account of Context Effects in Attitude Surveys. In Social Information Processing and Survey Methodology. H.J. Hippler et al. (eds). New York: Springer Verlag, 123–148.
Sudman, S. and Bradburn, N.M. (1974). Response Effects in Surveys. Chicago: Aldine.
Tourangeau, R (1984). Cognitive Science and Survey Methods: A Cognitive Perspective. In Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. T. Jabine et al. (eds). Washington DC: National Academy Press, 73–100.
Turner,C.F., Forsyth, B.H., O'Reilly, J.M., Cooley, P.C., Smith, T.K., Rogers, S.M., and Miller, H.G. (1998). Automated Self-interviewing and the Survey Measurement of Sensitive Behaviors. In Computer Assisted Survey Information Collection. M.P. Couper, R.P. Baker., J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls II, and J.M. O'Reilley (eds). New York, Wiley, 455–473.
Vermunt, J. (1996). Causal Log-linear Modeling with Latent Variables and Missing Data. In Analysis of Change. Advanced Techniques in Panel Data Analysis. U. Engel and J. Reinecke (eds). New York: De Gruyter.
Wothke, W. (2000). Longitudinal and Multigroup Modelling with Missing Data. In Modeling Longitudinal and Multilevel Data. T.D. Little, K.U. Schnabel, and J. Baumert (eds). Mahwah, NJ: Lawrence Erlbaum Associates, 219–240.

Received February 2000
Revised October 2002