CHAPTER 1

Survey Error Evaluation

1.1 SURVEY ERROR

1.1.1 An Overview of Surveys

This book focuses primarily on the errors in data collected in sample surveys and how to evaluate them. A natural place to start is to define the term "survey." The American Statistical Association's Section on Survey Research Methods has produced a series of 10 short pamphlets under the rubric *What Is a Survey?* (Scheuren 1999). That series defines a survey as a *method* of gathering information from a *sample* of objects (or *units*) that constitute a *population*. Typically, a survey involves a questionnaire of some type that is completed by either an informant (referred to as the *respondent*), an interviewer, an observer, or other agent acting on behalf of the survey organization or sponsor. The population of units can be individuals such as householders, teachers, physicians, or laborers or other entities such as businesses, schools, farms, or institutions. In some cases, the units can even be events such as hospitalizations, accidents, investigations, or incarcerations. Essentially any object of interest can form the population.

In a broad sense, surveys also include censuses because the primary distinction is just the fraction of the sample to be surveyed. A survey is confined to only a sample or subset of the population. Usually, only a small fraction of the population members is selected for the sample. In a *census*, every unit in the population is selected. Therefore, much of what we say about surveys also applies to censuses.

If the survey sample is selected randomly (i.e., by a probability mechanism giving known, nonzero probabilities of selection to each population member), valid statistical statements regarding the parameters of the population can be made. For example, suppose that a government agency wants to estimate the proportion of 2-year-old children in the country that has been vaccinated against infectious diseases (polio, diphtheria, etc.). A randomly selected sample

Latent Class Analysis of Survey Error By Paul P. Biemer

of 1000 children in this age group is drawn and their caregivers are interviewed. From these data, it is possible to determine the proportion of children who are vaccinated within some specified *margin of error* for the estimate. Sampling theory [see, e.g., Cochran (1977) or more recently, Levy and Lemeshow (2008)] provides specific methods for estimating margins of error and testing hypotheses about the population parameters.

Surveys may be cross-sectional or longitudinal. *Cross-sectional* surveys provide a "snapshot" of the population at one point in time. The products of cross-sectional surveys are typically descriptive statistics that capture distributions of the population for characteristics of interest, including health, education criminal justice, economics, and environmental variables. Cross-sectional surveys may occur only once or may be repeated at some regular interval (e.g., annually). As an example, the National Heath Interview Survey (Centers for Disease Control and Prevention 2009) is conducted monthly and collects important data on the health characteristics of the US population.

Longitudinal or *panel* surveys are repeating surveys where at least some of the same sample units are interviewed at different points in time. By taking similar measurements on the same units at different points in time, investigators can more precisely estimated changes in population parameters as well as individual characteristics. A *fixed panel* (or *cohort*) survey interviews the entire sample repeatedly usually over some significant period of time such as 2 or more years. As an example, the Panel Study of Income Dynamics (Hill 1991) has been collecting income data on the same 4800 families (as well as families spawned from these) since 1968.

A *rotating panel survey* is a type of longitudinal survey where part of the sample is replaced at regular intervals while the remainder of the sample is carried forward for additional interviewing. This design retains many of the advantages of a fixed panel design for estimating change while reducing the burden and possible conditioning effects on sample units caused by repeatedly interviewing them many times. An example is the US Current Population Survey (CPS) (US Census Bureau 2006), which is a monthly household survey for measuring the month-to-month and year-to-year changes in labor force participation rates. The CPS uses a somewhat complex rotating panel design, where each month about one-eighth of the sample is replaced by new households. In this way, households are interviewed a maximum of 8 times before they are *rotated* out of the sample.

Finally, a *split-panel* survey is a type of longitudinal survey that combines the features of a repeated cross-sectional survey with a fixed panel survey design. The sample is divided into two subsamples: one that is treated as a repeated cross-sectional survey and the other that follows a rotating panel design. An example of a split-sample design is the American National Election Studies [American National Election Studies (ANES), 2008]. Figure 1.1 compares these four survey designs.

As this book will explain, methods for evaluating the error in surveys may differ depending on the type of survey. Many of the methods discussed can be

	Time 1	Time 2	Time 3	Time 4
Repeated Cross-Sectional				
Sample 1	X			
Sample 2		X		
Sample 3			X	
Fixed Panel				
Sample 1	X	X	X	X
Rotating Panel				
Sample 1	X	X		
Sample 2		X	X	
Sample 3			X	X
Split-Panel				
Sample 1-A	X			
Sample 1-B	X	X		
Sample 2-A		X		
Sample 2-B		X	X	
Sample 3-A			X	
Sample 3-B			X	X

Figure 1.1 Reinterview patterns for four survey types.

applied to any survey while others are appropriate only for longitudinal surveys. The next section provides some background on the problem of survey error and its effects on survey quality.

1.1.2 Survey Quality and Accuracy and Total Survey Error

The terms *survey quality*, *survey data quality*, *accuracy*, *bias*, *variance*, *total survey error*, *measurement validity*, and *reliability* are encountered quite often in the survey error literature. Unfortunately, their definitions are often unspecified or inconsistent from study to study, which has led to some confusion in the field. In this section, we provide definitions of these terms that are reasonably consistent with conventional use, beginning with perhaps the most ambiguous term: survey quality.

Because of its subjective nature, *survey quality* is a vague concept. To some data producers, survey quality might mean *data* quality: large sample size, a high response rate, error-free responses, and very little missing data. Statisticians, in particular, might rate such a survey highly on some quality scale. Data users, on the other hand, might still complain that the data were not timely or accessible, that the documentation of the data files is confusing and incomplete, or that the questionnaire omitted many relevant areas of inquiry that are essential for research in their chosen field. From the user's perspective, the survey exhibits very poor quality.

These different points of view suggest that survey quality is a very complex, multidimensional concept. Juran and Gryna (1980) proposed a simple definition of quality that can be appropriately applied to surveys, namely, the quality of a product is its "fitness for use." But, as Juran and Gryna explain, this definition is deceptively simple because there are really two facets of quality: (1) freedom from deficiencies and (2) responsiveness to customers' needs. For survey work, facet 1 might be translated as error-free data, data accuracy, or high data quality, while facet 2 might be translated as providing product features that result in high user satisfaction. The latter might include data accessibility and clarity, timely data delivery, collection of relevant information, and use of coherent and conventional concepts.

When applied to statistical products, the definition "fitness for use" has another limitation in that it implies a single use or purpose. Surveys are usually designed for multiple objectives among many data users. A variable in a survey may be used in many different ways, depending on the goals of the data analyst. For some uses, timeliness may be paramount. For other uses, timeliness is desirable, but *comparability* (i.e., ensuring that the results can be compared unambiguously to prior data releases from the same survey) may be more critical.

In the mid-1970s, a few government statistical offices began to develop definitions for survey quality that explicitly took into account the multidimensionality of the concept [see, e.g., Lyberg et al. (1977) or, more recently, Fellegi (1996)]. This set of definitions has been referred to as a *survey quality framework*. As an example, the quality framework used by Statistics Canada includes these seven quality dimensions: relevance, accuracy, timeliness, accessibility, interpretability, comparability, and coherence. Formal and accepted definitions of these concepts can be found at Statistics Canada (2006). Eurostat has also adopted a similar quality framework [see, e.g., Eurostat (2003)].

Given this multidimensional conceptulization of quality, a natural question is quality to be maximized in a survey? One might conceptualize a one-dimensional indicator that combines these seven dimensions into an overall survey quality indicator. Then the indicator could be evaluated for various designs and the survey design maximizing this quantity could be selected. However, this approach oversimplifies the complexity of the problem since there is no appropriate way for combining the diverse dimensions of survey quality. Rather, quality reports or quality declarations providing information on each dimension have been used to summarize survey quality. A quality report might include a description of the strengths and weaknesses of a survey organized by quality dimension, with emphasis on sampling errors, nonsampling errors,[1] key release dates for user data files, forms of dissemination,

[1]Nonsampling errors are discussed in more detail in Section 1.1.3. Nonsampling errors, which are inevitable in survey data collection, arise from many sources, including interviewers, respondents, data processors, other survey personnel, and operations. As the term implies, they are essentially all errors in a survey apart from sampling errors.

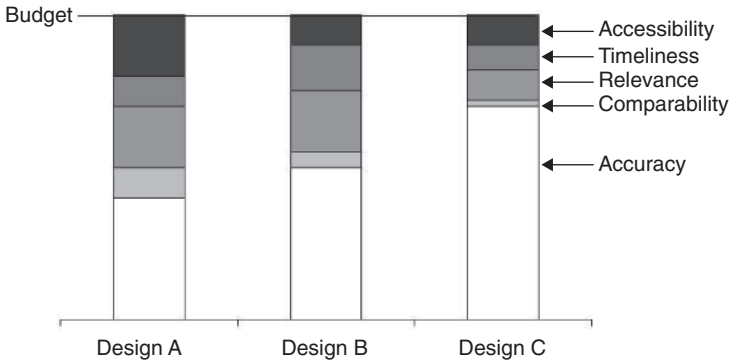

Figure 1.2 Comparisons of three cost-equivalent survey designs. (*Source*: Biemer 2010).

availability, and contents of documentation, as well as special features of the survey approach that may be of importance to most users. A number of surveys have produced extended versions of such reports, called *quality profiles*. A quality profile is a document that provides a comprehensive picture of the quality of a survey, addressing each potential source of error. Quality profiles have been developed for a number of US surveys, including the Current Population Survey (CPS) (Brooks and Bailar 1978), the Survey of Income and Program Participation (Jabine et al. 1990), US Schools and Staffing Survey (Kalton et al. 2000), American Housing Survey (Chakrabarty and Torres 1996), and the US Residential Energy Consumption Survey (Energy Information Administration 1996). Kasprzyk and Kalton (2001) review the use of quality profiles in US statistical agencies and discuss their strengths and weaknesses for survey improvement and quality declaration purposes.

Note that *data quality* or *accuracy* is not synonymous with *survey quality*. Good survey quality is the result of optimally balancing all quality dimensions to suit the specific needs of the primary data users. As an example, if producing timely data is of paramount importance, accuracy may have to be compromised to some extent. Likewise, if a high level of accuracy is needed, temporal comparability may have to be sacrificed to take advantage of the latest and much improved methodologies and technologies. On the other hand, *data quality* refers to the amount of error in the data. As such, it focuses on just one quality dimension—accuracy.

To illustrate this balancing process, Figure 1.2 shows three cost-equivalent survey designs, each with a different mix of five quality dimensions: accessibility, timeliness, relevance, comparability, and accuracy. The shading of the bars in the graph represents the proportion of the survey budget that is to be allocated for each quality dimension. For example, design C allocates about two-thirds of the budget to achieve data accuracy while design A allocates less to accuracy so that more resources can be devoted to the other four dimensions. Design B represents somewhat of a compromise between designs A and C.

Determining the best design allocation depends on the purpose of the survey and how the data will ultimately be used. If a very high level of accuracy is required (e.g., a larger sample or a reduction of nonsampling errors), design C is preferred. However, if users are willing to sacrifice data quality for the sake of greater accessibility, relevance, and comparability, then design A may be preferred. Since each design has its own strengths and weaknesses, the best one will have a mix of quality attributes that is most appropriate for the most important purposes or the majority of data users.

Total survey error refers to the totality of error that can arise in the design, collection, processing, and analysis of survey data. The concept dates back to the early 1940s, although it has been revised and refined by a many authors over the years. Deming (1944), in one of the earliest works, describes "13 factors that affect the usefulness of surveys." These factors include sampling errors as well as nonsampling errors: the other factors that will cause an estimate to differ from the population parameter it is intended to estimate. Prior to Deming's work, not much attention was being paid to nonsampling errors, and, in fact, textbooks on survey sampling seldom mentioned them. Indeed, classical sampling theory (Neyman 1934) assumes that survey data are error-free except for sampling error. The term *total survey error* originated with an edited volume of the same name (Andersen et al. 1979).

Optimal survey design is the process of minimizing the total survey error subject to cost constraints [see, e.g., Groves (1989) and Fellegi and Sunter (1974)]. Biemer and Lyberg (2003) extended this idea to include other quality dimensions (timeliness, accessibility, comparability, etc.) in addition to accuracy. They advocate an approach that treats the other quality dimensions as additional constraints to be met as total survey error is minimized (or equivalently, accuracy is maximized). For example, if the appropriate balance of the quality dimensions is as depicted by design B in Figure 1.2, then the optimal survey design is one that minimizes total survey error within that fraction of the budget allocated to achieving high data accuracy represented by the unshaded area of the bar. As an example, in the case of design B, the budget available for optimizing accuracy is approximately 50% of the total survey budget. The optimal design is one that maximizes accuracy within this budget allocation while satisfying the requirements established for the other quality dimensions shown in the figure.

Mean-Squared Error (MSE)

The prior discussion can be summarized by stating that the central goal of survey design should be to minimize total survey error subject to constraints on costs while accommodating other user-specified quality dimensions. *Survey methodology*, as a field of study, aims to accomplish this goal. General textbooks on survey methodology include those by Groves (1989), Biemer and Lyberg (2003), Groves et al. (2009), and Dillman et al. (2008), as well as a number of edited volumes. The current book focuses on one important facet of survey methodology—the evaluation of survey error, particularly measure-

ment error. The book focuses on the current best methods for assessing the accuracy of survey estimates subject to classification error. A key concept in the survey methods literature is the *mean squared error* (MSE), which is a measure of the accuracy of an estimate. The next few paragraphs describe this concept.

Let $\hat{\mu}$ denote an estimate of the population parameter μ based on sample survey data. *Survey error* may be defined as the difference between the estimate and the parameter that it is intended to estimate:

$$\text{Survey error} = \hat{\mu} - \mu \tag{1.1}$$

There are many reasons why $\hat{\mu}$ and μ may disagree and, consequently, the survey error will not be zero. One obvious reason is that the estimator of μ is based upon a sample and, depending on the specific sample selected, $\hat{\mu}$ will deviate from μ, sometimes considerably so, especially for small samples. However, even in very large samples, the difference can be considerable due to *nonsampling errors*, meaning errors in an estimate that arise from all sources other than sampling error. The survey responses themselves may be in error because of ambiguous question wording, respondent errors, interviewer influences, and other sources. In addition, there may be missing data due to nonresponding sample members (referred to as *unit nonresponse*) or when respondents do not answer certain questions (referred to as *item nonresponse*). Data processing can also introduce errors. All these errors can cause $\hat{\mu}$ and μ to differ even when there is no sampling error as in a complete census.

In the survey methods literature, the preferred measure of *total survey error of an estimate* (i.e., the combination of sampling and nonsampling error sources) is the MSE, defined as

$$\text{MSE}(\hat{\mu}) = E(\hat{\mu} - \mu)^2 \tag{1.2}$$

which can be rewritten as

$$\text{MSE}(\hat{\mu}) = B^2 + Var(\hat{\mu}) \tag{1.3}$$

where $B = E(\hat{\mu} - \mu)$ is the bias of the estimator and $Var(\hat{\mu})$ is the variance. In these expressions, *expected value* is broadly defined with respect to the sample design as well as the various random processes that generate nonsampling errors. Optimal survey design attempts to minimize (1.3) given the budget, schedule, and other constraints specified by the survey design. This is a very challenging task. It is facilitated, first and foremost, by some assessment of the magnitude of the MSE for at least a few key survey characteristics.

The preferred approach to evaluating survey error is to first decompose the total error into components associated with the various sources of error in a survey. Then each error source can be evaluated separately in a "divide and conquer" fashion. Some sources may be ignored while others may be targeted

in special evaluation studies. Biemer and Lyberg (2003, Chapter 2) suggest a mutually exclusive and exhaustive list of error sources applicable to most surveys. The list includes sampling error and five nonsampling error sources: specification error, measurement error, nonresponse error, frame error, and data processing error. These errors sources are briefly described next.

1.1.3 Nonsampling Error

The evaluation of sampling error is considered a best practice in survey research. Nonsampling errors are rarely fully evaluated in surveys, although many examples of evaluations focus on one or perhaps two error sources. In this section, we consider the five sources of nonsampling error in more detail and then discuss some methods that have been used in their evaluation.

Specification Error

A *specification error* arises when the concept implied by the survey question and the concept that should have been measured in the survey differ.[2] When this occurs, the wrong parameter is estimated by the survey and, thus, inferences based on the estimate are likely to be erroneous. Specification error is often caused by poor communication between the researcher (or subject matter expert) and the questionnaire designer. This concept is closely related to the concept of *construct validity* in psychometric literature [see, e.g., Nunnally and Bernstein (1994)] and *relevance* in official statistics [see, e.g., Dalenius (1985)].

Specification errors are particularly common in surveys of business establishments and organizations where many terms that have precise meanings to accountants are misspecified or defined incorrectly by the questionnaire designers. Examples are terms such as *revenue*, *asset*, *liability*, *gross service fees*, and *information services*, which have different meanings in different contexts. Such specialized terms should be clearly defined in surveys to avoid specification error.

As an example, consider the measurement of unemployment in the US Current Population Survey (CPS). The US Bureau of Labor Statistics (BLS) considers the unemployed population as comprising two types of persons: those who are "looking for work" and those who are "on layoff." Persons on layoff are defined as those who are separated from a job and await a recall to return to that job. Persons who are "looking for work" are the unemployed who are not on layoff and who are pursuing certain specified activities to find employment. Distinguishing between these two groups is important for labor economists. Prior to 1994, the CPS questionnaire did not consider or collect information as to whether a person classified as "laid off" expected to be

[2]This usage of the term should not be confused with the econometric term *model specification error* or *misspecification*. The latter error arises through omission of important variables from a statistical model [see, e.g., Ramsey (1969)] or failures of model assumptions to hold.

recalled to work at some point in the future. Rather, respondents were simply asked "Were you on layoff from a job?" This question was later determined to be problematic because, to many people, a "layoff" could mean *permanent termination* from the job rather than the temporary loss of work as the BLS economists defined the term.

In 1994, the BLS redesigned the labor force status questions, as part of that redesign, attempted to clarify the concept of layoff in the questionnaire. The revised questions now ask, "Has your employer given you a date to return to work?" and "Could you have returned to work if you had been recalled?" These questions brought the concept of "on layoff" in line with the specification being used by BLS economists. Specification errors can be quite difficult to detect without the help of subject matter experts who are intimately familiar with the survey concepts and how they will ultimately be used in data analyses, because questions may be well-worded while still completely missing essential elements of the variable to be measured.

Biemer and Lyberg (2003, p. 39) provide another example of specification error from the Agriculture Land Values Survey (ALVS) conducted by the US National Agricultural Statistics Service. The ALVS asked farm operators to provide the market value for a specific tract of land that was randomly selected within the boundaries of the farm. Unfortunately, the concepts that were essential for the valid valuation of agricultural land were not accurately stated in the survey—a problem that came to light only after economists at the Economic Research Service (ERS) were consulted regarding the true purpose of the questions. These subject matter experts pointed out that their models required a value that did not include capital improvements such as irrigation equipment, storage facilities, and dwellings. Because the survey question did not exclude capital improvements, the survey specification of agricultural land value was inconsistent with the way the ERS economists were using the data.

Measurement Error

Whereas *total survey error* is defined for a statistic (or estimator), *measurement error* is defined for an observation. Let μ_i denote the true value of some characteristic measured in a survey for a unit i, and let y_i denote the corresponding survey measurement of μ_i. Then

$$\text{Measurement error} = y_i - \mu_i \tag{1.4}$$

that is, the difference between the survey measurement and the true value of the characteristic. Measurement error has been studied extensively and is often reported in the survey methods literature [for an extensive review, see Biemer and Lyberg (2003, Chapters 4–6)]. For many surveys, measurement error can also be the most damaging source of error. It includes errors arising from respondents, interviewers, and survey questions. Respondents may either deliberately or otherwise provide incorrect information in response to questions. Interviewers can cause errors in a number of ways. They may, by their

appearance or comments, influence responses; they may record responses incorrectly, or otherwise fail to comply with prescribed survey procedures; and, in some cases, they may deliberately falsify data. The questionnaire can be a major source of error if it is poorly designed. Ambiguous questions, confusing instructions, and easily misunderstood terms are examples of questionnaire problems that can lead to measurement error.

Measurement errors can also arise from the information systems that respondents may draw on to formulate their responses. For example, a farm operator or business owner may consult records that may be in error and, thus, cause an error in the reported data. It is also well known (Biemer and Lyberg 2003, Chapter 6) that the mode of data collection can affect measurement error. As an example, mode comparison studies (Biemer 1988; de Leeuw and van der Zouwen 1988; Groves 1989) have found that data collected by telephone interviewing are, in some cases, less accurate than the same information collected by face-to-face interviewing. Finally, the setting or environment within which the survey is conducted can also contribute to measurement error. When collecting data on sensitive topics such as drug use, sexual behavior, or fertility, the interviewer may find that a private setting is more conducive to obtaining accurate responses than one in which other members of the household are present. In establishment surveys, topics such as land use, financial loss and gain, environmental waste treatment, and resource allocation can also be sensitive. In these cases, assurances of confidentiality may reduce measurement errors that result from intentional misreporting. Biemer et al. (1991) provides a comprehensive review of measurement error in surveys.

Frame Error

Frame error arises in the process for constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. The sampling frame is defined as a list of population members or some other mechanism used for drawing the sample. Ideally, the frame would contain every member of the population with no duplicates. Also, units that are not part of the population would not be on the frame. Likewise, information on the frame that is used in the sample selection process should be accurate and up to date. Unfortunately, sampling frames rarely satisfy these ideals, often resulting in various types of frame errors.

There are essentially three types of sampling frames: area frames, list frames, and implicit frames. *Area frames* are typically used for agricultural and household surveys. An area frame is constructed by first dividing an area to be sampled (say, a state) into smaller areas (such as counties, census tracts, or blocks). A random sample of these smaller areas is drawn and a *counting and listing* operation is implemented in the selected areas to enumerate all the ultimate sampling units. For household surveys, the counting and listing operation is intended to identify and list every dwelling unit in the sampled smaller areas. Following the listing process, dwelling units may be sampled according to any appropriate randomization scheme. The process is

similar for agricultural surveys, except rather than a dwelling unit, the ultimate sampling unit may be a farm or land parcel.

The omission of eligible population units from the frame (referred to as *noncoverage error*) can be a problem with area samples, primarily as a result of errors made during the counting–listing phase. Enumerators in the field may miss some dwelling units that are hidden from view or are mistaken as part of other dwelling units (e.g., garages that have been converted to apartments). Boundary units may be erroneously excluded or included because of inaccurate maps or enumerator error. Boundary units can also be a source of duplication error if they are included for areal units on both sides of the boundary.

More recent research has considered the use of *list frames* for selecting household samples [see, e.g., O'Muircheartaigh et al., (2007), Dohrmann et al. (2007), and Iannacchione et al. (2007)]. One such list is the US Postal Service *delivery sequence file* (DSF). This frame contains all the delivery point addresses serviced by the US Postal Service. Because sampling proceeds directly from this list, a counting–listing operation is not needed, saving considerable cost. Noncoverage error may be an important issue in the use of the DSF, particularly in rural areas [see, e.g., Iannacchione et al. (2003)]. Methods for reducing the noncoverage errors, such as the *half-open interval method* [see, e.g., Groves et al. 2009)] have met with varying success (O'Muircheartaigh et al., 2007). List frames are also commonly used for sampling special populations such as teachers, physicians, and other professionals. Establishment surveys make extensive use of list frames drawn from establishment lists purchased from commercial vendors.

A sampling frame may not be a physical list, but rather an implicit list as in the case of random-digit dialing (RDD) sampling. For RDD sampling, the frame is implied by the mechanism generating the random numbers. Frame construction may begin by first identifying all telephone exchanges (e.g., in the United States and Canada, the area code plus the 3-digit prefix) that contain at least one residential number. The implied frame is then all 10-digit telephone numbers that can be formed using these exchanges, although the numbers in the sample are the only telephone numbers actually generated and eventually dialed. *Intercept sampling* may also use an implicit sampling frame. In intercept sampling, a systematic sample of units is selected as they are encountered during the interviewing process; examples where an explicit list of population units is not available include persons in a shopping mall or visitors to a website.

To ensure that samples represent the entire population, every person, farm operator, household, establishment, or other element in the population should be listed on the frame. Ineligible units should be identified and removed from the sample as they are selected. Further, to weight the responses using the appropriate probabilities of selection, the number of times that each element is listed on the frame should also be known, at least for the sampled units. To the extent that these requirements fail, frame errors occur.

Errors can occur when a frame is constructed. Population elements may be omitted or duplicated an unknown number of times. There may be elements on the frame that should not be included (e.g., in a farm survey, businesses that are not farms). Erroneous omissions often occur when the cost of creating a complete frame is too high. We may be well aware that the sampling frame for the survey is missing some units but the cost of completing the frame is quite high. If the number of missing population members is small, then it may not be worth the cost to provide a complete frame. Duplications on a frame are a common problem when the frame combines a number of lists. For the same reason, erroneous inclusions on the frame usually occur because the available information about each frame member is not adequate to determine which units are members of the population and which are not. Given these frame imperfections, the population represented by the frame does not always coincide with the population of interest in the survey. The former population is referred to as the *frame population* and the latter as the *target population*.

Nonresponse Error

Nonresponse error is a fairly general source of error encompassing both unit and item nonresponse. Unit nonresponse occurs when a sampled unit (e.g., a household, farm, school or establishment) does not respond to any part of a questionnaire, such as a household that refuses to participate in a face-to-face survey, a mail survey questionnaire that is never returned, or an eligible sample member who refuses or whose telephone is never answered. Item nonresponse error occurs when the questionnaire is only partially completed because an interview was prematurely terminated or some items that should have been answered were skipped or left blank. For example, income questions are typically subject to a high level of item nonresponse from respondent refusals.

For open-ended questions, even when a response is provided, nonresponse may occur if the response is unusable or inadequate. As an example, a common open-ended question in socioeconomic surveys is "What is your occupation?" A respondent may provide some information about his or her occupation, but perhaps not enough to allow an occupation and industry coder to assign an occupation code number during the data-processing stage.

Data-Processing Error

The final source of nonsampling error is data processing. Data-processing error includes errors in editing, data entering, coding, weighting, and tabulating of the survey data. As an example of editing error, suppose that a data editor is instructed to call the respondent back to verify the value of some budget line item whenever the value of the item exceeds a specified limit. In some cases, the editor may fail to apply this rule correctly, thus generating errors in the data.

For open-ended items that are subsequently coded, coding error is another type of data-processing error. The coders may make mistakes or deviate from prescribed procedures. The system for assigning the code numbers—for

variables such as place of work, occupation, industry in which the respondent is employed, and field of study for college students—may itself be ambiguous and prone to error. As a result, code numbers may be inconsistently and inappropriately assigned, resulting in significant levels of coding error.

The survey weights that statistically compensate for unequal selection probabilities, nonresponse error, and frame coverage errors may be calculated erroneously, or there may be programming errors in the estimation software that computes the weights. Errors in the tabulation software may also affect the final data tables. For example, a spreadsheet used to compute the estimates may contain a cell-reference error that goes undetected. As a result, the weights are applied incorrectly and the survey estimates are in error.

Decomposing Total Survey Error

While each source of error can increase the total error, some pose a much greater risk for bias and/or variance than others. Biemer and Lyberg (2003) provide a rough assessment of the risk of *variable* and *systematic* errors for each error source based on a synthesis of the nonsampling error literature. They define variable errors as errors that are distributed with zero mean and nonzero variance. Systematic errors are defined as errors having nonzero mean and zero or trivially small variance. As an example, in a survey using acceptable probability sampling methods, the sampling error distribution has zero mean and there is no risk of bias. Of course, sampling variance is an unavoidable consequence of sampling.

Specification error contributes systematic errors because measuring the wrong concept in a survey causes the estimate to consistently differ from the true concept. Nonresponse error typically poses a greater risk to systematic error, although nonresponse adjustment methods such as imputation and weighting can contribute importantly to the variance, especially when the nonresponse rates are high. Frame errors, particularly noncoverage errors, are viewed primarily as a source of systematic error, although, as with nonresponse, coverage adjustments can also increase the variance. Measurement errors (emanating primarily from interviewers, respondents, and the questions themselves) can pose a serious risk for both systematic and variable errors, as detailed in the later chapters. Finally, data-processing operations involving human operators, such as coding and editing, can add both systematic and variable errors to the total error.

Using these risk classifications, Biemer and Lyberg posit an expanded version of the MSE that is applicable to estimates of means, totals, and proportions and that includes specific terms for each major source of error. In their formulation, the B^2 component is expanded to include bias components for all the sources of error having a high risk of systematic error: specification bias, B_{SPEC}; frame bias, B_{FR}; nonresponse bias, B_{NR}; measurement bias, B_{MEAS}; and data-processing bias, B_{DP}. These components sum together to produce the total bias component, B:

$$B^2 = (B_{\text{SPEC}} + B_{\text{FR}} + B_{\text{NR}} + B_{\text{MEAS}} + B_{\text{DP}})^2 \tag{1.5}$$

Likewise, the variance of an estimator can be decomposed into components associated with the major sources of variable error: sampling variance (Var_{SAMP}), measurement error variance (Var_{MEAS}), and data-processing variance (Var_{DP}). Combining these components, an expanded version of the MSE formula showing components for all the major sources of bias and variance can be written as

$$\text{MSE}(\hat{\mu}) = (B_{SPEC} + B_{FR} + B_{NR} + B_{MEAS} + B_{DP})^2 + Var_{SAMP} + Var_{MEAS} + Var_{DP} \tag{1.6}$$

Although complex, this expression for the MSE is still oversimplified because it ignores some interaction terms that may be important for some applications. The article by Fellegi (1964) contains an excellent discussion of these interaction terms. The use of this "divide and conquer" approach greatly simplifies the task of evaluating total survey error because scarce evaluation resources can be targeted to the error source(s) with the greatest effect on survey accuracy, provided the effect can be assessed.

Another way to classify the nonsampling error sources is by errors associated with missing data, or *errors of nonobservation* (viz., nonresponse and frame errors) and those that affect the content of the observations (viz., specification, measurement, and data processing). The focus of this book is on the latter, which are sometimes referred to as *content* errors[3] but more often as *measurement errors*.

1.2 EVALUATING THE MEAN-SQUARED ERROR

As discussed in this section, estimating the MSE or any of its components, often requires the collection of additional data, complex methods of analysis, or both. These are costly and may consume resources that could otherwise be put toward the data collection effort to improve survey quality. Why then should survey organizations bear this expense and allocate resources for survey evaluation? This question is addressed in the next section.

1.2.1 Purposes of MSE Evaluation

Historically, total survey error components have been estimated for at least four purposes:

To compare the accuracy of data from alternative modes of data collection or estimation methods

[3]Groves (1989) refers to these errors as *errors of observation*. We prefer our terminology to reflect those errors (such as specification and data processing errors) that occur either before or after data are observed, yet still affect the content of the data.

To optimize the allocation of resources for the survey design

To reduce the nonsampling error contributed by specific survey processes

To provide information to data users regarding the quality of the data or the reported estimates

The first is one of the most common uses of total survey error evaluations. As an example, a survey methodologist may wish to compare the accuracy of health data collected by mail and by telephone. Because mail is usually the cheaper mode, a larger mail survey could be afforded that would reduce sampling error, but it is possible that the total survey error would increase because of nonsampling errors. To compare the two modes, a mode comparison study could be conducted using a *split-ballot design* in which half the sample is collected by telephone and the other half is collected by mail. But while the split-ballot approach is sufficient for determining whether the two modes give differing results, it is seldom sufficient for determining the more accurate mode. For evaluating accuracy, the MSEs of the estimates from each mode must be evaluated and compared (Biemer 1988).

In many cases, it is also important to determine whether the differences in MSEs stem from bias or variance components and which components contribute most to the differences. For example, if the response rates differ considerably for the two modes, it may be tempting to attribute the difference in accuracy for some key characteristics to nonresponse bias. For other characteristics, the real culprit may be measurement bias or variance. Knowledge of which MSE components are most responsible for the differences in MSE will provide clues as to how the differences can be minimized or eliminated.

Another purpose of survey error evaluation is design optimization. For this purpose, the survey designer would like to know how much of the error is contributed by each major error source. Then resources could be allocated for mitigating the sources of the largest errors. An alternative is to use expert judgment or intuition for allocating resources, but the risk with that approach is that the allocation may be far from optimal for minimizing total survey error.

An example of optimal survey design is Dillman's (2007) *tailored design method* for mail surveys. This methodology has been developed by combining the results of many experiments across many surveys and on a very wide range of topics. Dillman and others have used meta-analysis and other techniques for integrating this vast collection of research results. In the process, they have identified the most essential factors of optimal design and the best combinations of sample design, questionnaire design, and implementation techniques for minimizing nonresponse and measurement error and reducing survey costs.

Another important reason for estimating the MSE is that information on the magnitudes of nonsampling error components contributed by specific survey operations is also useful for identifying where improvements are

needed for ongoing survey operations. As an example, a study of errors made by interviewers may determine that interviewers contribute considerably to the total variance as well as the bias of estimates. Additional testing and experimentation could be done to identify the root causes of the error focusing on the interviewer variance component. Further study may reveal that improvements are needed in interviewer training, the questionnaire, interviewer monitoring or supervision, or other areas. Over the years, studies of the components of total survey error have led to many improvements in survey accuracy. For example, studies of enumerator variance in the 1950 US census led to the decision to adopt a self-enumeration census methodology (Eckler 1972, p. 105).

Finally, the estimation of the total MSE and its individual components can provide data users with objective information on the relative importance of different errors, which can aid their understanding of the limitations of the data. Measures of nonsampling error indicating excellent or very good data quality create high user confidence in the quality of the data, while measures that imply only fair to poor data quality can serve as a warning to users to proceed with caution in interpretation of the data.

As an example, reports on survey quality often contain estimates of nonresponse bias for the key estimates produced from the survey data. This information is quite informative for assessing the accuracy of the estimates and whether nonsampling error should be a concern for interpreting the research findings. Likewise, an analysis of measurement error could be useful for explaining why an analysis failed to replicate findings in the literature or why unexpected and inexplicable relationships among the variables were found.

To understand the causes of nonsampling error and develop strategies for its prevention, the errors must be measured often. Continuous quality improvement requires current knowledge of which error sources are the most problematic so that scarce survey resources can be most effectively allocated. For some error components, this might involve interviewing a small representative sample of the target population using *cognitive interviewing methods* [see, e.g., Forsyth and Lessler (1991), Tourangeau et al. (2000), and Willis (2005)] rather than a large study aimed at estimating a bias component. However, small-scale laboratory investigations used in conjunction with large-scale error component evaluation studies may be ideal for most purposes. Evaluation studies aimed at describing the effect of alternate design choices on total survey error are also extremely important, because without them total survey design optimization is not possible.

1.2.2 Effects of Nonsampling Errors on Analysis

One of the primary motivations for wanting to minimize survey errors in surveys is their potentially devasting effects on all types of data analysis. Cochran (1968), Fuller (1987), and more recently Biemer and Trewin (1997)

consider the effects of nonsampling errors on a range of estimators and analysis methods including the estimation of means, quantiles and their standard errors, correlations, regression coefficients [including analysis of covariance and analysis of variance (ANOVA)], goodness of fit, and association test statistics in contingency table analysis. As they show, the damaging effects of nonsampling errors on estimation and inference can be quite severe. Some evidence of this damage is summarized here; however, more details regarding these effects is provided in later chapters.

As noted above, variable errors may not be biasing for estimators of means, totals, and proportions. Variable errors will inflate the variance of these estimators. Biemer and Trewin show the inflation factor to be the reciprocal of the reliability *reliability ratio* for the measurements.[4] The reliability ratio, denoted by R, may be defined roughly as the proportion of the total variance of an observation that is not measurement error variance. They further show that the usual estimator of the standard error will be unbiased when nonsampling errors are variable only. In other words, although the standard error is inflated, the usual estimator or the standard error correctly reflects the increase. An exception is when the nonsampling errors are correlated in ways not reflected in the survey design.

Correlated errors occur when survey operators (interviewers are a prime example) contribute errors to the observations that vary in magnitude across the operators [see, e.g., Kish (1962), Fellegi (1964) Fellegi and Sunter (1974), Biemer and Stokes (1991), and Biemer and Lyberg (2003)]. For example, suppose that interviewers are asked to estimate the average value of housing in the neighborhoods where their sampling units are located. Some interviewers who are not current on the housing values may consistently undersestimate these values for the units in their work assignments. Others may overestimate these values for other reasons. Similarly, interviewers, by the way they dress or act in an interview, may influence responses in a systematic way. These systematic errors may be regarded as random effects that covary positively across the units in an interviewer's assignment. Thus, two observations on different units in the same interviewer's assignment may be positively correlated as a result of these *correlated interviewer errors*.

Similar to a "clustering effect" in two-stage sampling [see, e.g., Kish (1965, pp. 257–259)], correlated interviwer (or, more genereally, operator) errors increase the variance of estimators by a multiplicative factor that is a function of average caseload size and the intra-cluster correlation due to operators. For interviewing, the *interviewer clustering effect* is given by

$$\delta_{\text{int}} = [1 + (m-1)\rho_{\text{int}}] \tag{1.7}$$

[4]The reliability ratio is similar to the *signal-to-noise* ratio in physics. It will be discussed in detail in the Chapter 2.

where m is the average size (in number of units) of the interviewers' caseloads and ρ_{int} is the intraclass correlation of the interviewers.[5] Note the similarity of (1.7) to the design effect defined for cluster sampling (Kish 1965). As an example, suppose that the interviewers for a survey interviewed an average of 50 households each. Suppose further that ρ_{int} is estimated to be 0.01. [Biemer and Lyberg (2003) describe the methodology for estimating ρ_{int} in some detail.] Then, according to (1.7), the variance of an estimator of the mean will be increased by $1 + (49)0.01 = 1.49$ as a result of correlated interviewer error. This is essentially equivalent to a reduction of sample size by two-thirds for estimating the mean of the item. In fact, the sample size n divided by (1.7) is often referred to as the *effective sample size*.

A number of authors [see, e.g., Hansen et al. (1964), Cochran (1968), and Biemer and Trewin (1997)] have shown that, although standard errors may be inflated by variable errors, the traditional estimators of standard errors are still unbiased as long as the errors are uncorrelated. This is somewhat of a "silver lining" on the variable error "cloud." However, as Hansen et al. (1951) showed, correlated errors will cause the usual standard error estimators to be negatively biased. The extent of this bias is roughly the same as the magnitude of the increase in standard error from correlated errors. From the illustration above for interviewers, this bias can be substantial even when the intrainterviewer correlation is quite small (i.e., 0.01 or less).

For regression and correlation coefficients, variable errors tend to bias coefficients toward 0 referred to as *attenuation* toward 0. Fuller (1987) shows that, in ordinary regression analysis, the expected value of the estimated regression coefficient is $R \times \beta$, where R is the reliability ratio associated with the predictor variable and β is the regression coefficient assuming error-free measurements. In addition, the *coefficient of determination* (i.e., the usual regression R^2 measure), which measures the proportion of variance in the dependent variable explained by the independent variables, cannot exceed the reliability of the dependent variable in the regression (Fuller 1987).

Fuller further showed that the usual Pearson product moment correlation coefficient between two variables x and y is shown to have expected value $\sqrt{R_x R_y}\,\rho_{xy}$, where R_x and R_y are the reliability ratios associated with x and y, respectively, and ρ_{xy} is the correlation coefficient when x and y are measured without error. When nonsampling errors are correlated, the regression and correlation coefficients can be unpredictably biased in either a positive or a negative direction.

Biemer and Trewin (1997) show that estimators of population quantiles are biased whether the errors are variable (uncorrelated or correlated) or systematic. Likewise, parameter estimates from ANOVA and analysis of covariance (ANACOV), as well as chi-square (χ^2) tests for goodness of fit and association are biased even in the benign case of uncorrelated variable errors.

[5]The formula is discussed further in Section 8.3.

These results underscore the importance of collecting survey data that are nearly error-free. Essentially all statistical inferences are biased in the presence of nonsampling error, except in the case of linear point estimation and uncorrelated variable errors. Methods are now available for correcting data analysis for the effects of uncorrelated nonsampling errors. For example, structural equation models (SEMs) [see, e.g., Bollen (1989)], errors-in-variables regression analysis [see, e.g., Fuller (1987)], and latent class analysis [see, e.g., Hagenaars and McCutcheon (2002) as well as this book] are techniques that have been developed to deal with uncorrelated error in statistical inference. Software packages that employ these techniques are widely available. However, methods for handling correlated variable errors and systematic errors in statistical analysis are not so common. Exceptions include methods for compensating for the effects of item nonresponse and frame undercoverage through the application of postsurvey weight adjustments (Biemer and Christ 2008). For item nonresponse, methods for imputing missing data have been developed as a way of reducing the bias in data analysis [see de Waal and Haziza (2009) for a recent overview of the literature]. However, these methods are only partially effective at best. Notwithstanding these advances, eliminating nonsampling error during the survey process is still the best approach for avoiding nonsampling error bias in estimation and data analysis.

1.2.3 Survey Error Evaluation Methods

Survey data quality evaluation is a critical branch of the field of survey research because without it, the process of survey quality improvement is essentially guesswork. The best practices for conducting surveys are based on the results of survey quality evaluations.

A number of methods are available for evaluating survey error. Some methods can be applied during the design and pretesting stages to guide the designers as they attempt to optimize survey design. Several methods can be applied concurrently with data collection and data processing to monitor the quality of the data and to alert survey managers when important errors enter the survey process. Other methods can be applied on completion of the survey to describe the magnitudes of the error components and provide valuable information to survey designers for improving future surveys. Thus, quality evaluation can be seen as a continuous process that is carried from the design stage to the postsurvey analysis stage.

Table 1.1 summarizes of some of the most frequently used methods for survey error evaluation. These methods are organized in the order in which they might be used in the survey process: the survey design stage, the pretesting stage, the survey data collection stage, or the postsurvey stage. For a description of these methods, see Biemer and Lyberg (2003). As shown in Table 1.1, cognitive methods and interviewer debriefings tend to be used in the design and pretesting stages to identify questionnaire problems. Supervisor observations of interviewers, quality control methods, and the methods for

Table 1.1 Some Evaluation Methods and Their Purposes

Stage of the Survey Process	Evaluation Method	Purpose
Design/pretest	Expert review	Identify problems with questionnaire layout, format, question wording, question order, and instructions
	Cognitive methods	
	Behavior coding	
	Cognitive interviewing	
	Debriefings	
	Interviewer group discussions	Evaluate one or more stages of the response process
	Respondent focus groups	Evaluate data collection and processing procedures
Pretest/survey	Observation	Evaluate interviewer performance
	Supervisor observation	
	Telephone monitoring	Identify questionnaire problems
	Recorded interviews	
	Data quality control	Refine data collection procedures
	Nonresponse reduction analysis	Refine nonresponse followup procedures
Postsurvey	Postsurvey analysis	Compare alternative methods of data collection
	Experimental design	
	Embedded repeated measures	Estimate one or more MSE components
	Internal consistency	
	External validation	Validate survey estimates
	Postsurvey data collection	Identify problem questions
	Reinterview surveys	
	Record check studies	
	Nonresponse followup studies	

reducing nonresponse tend to be used during the survey. Postsurvey methods might include experimental design, reinterviews, and other methods aimed at estimating one or more components of the MSE. These categorizations are not definitive. For example, experimental design methods can also be used at the design and pretest stages to choose among competing methods. Likewise, interviewer and respondent debriefings can be used postsurvey to evaluate the effectiveness of various survey approaches.

For many surveys, data quality evaluation is limited to just a few activities, for instance, pretesting the questionnaire and data collection quality control monitoring procedures such as response rate monitoring. There may be additional quality control procedures for data processing operations such as data

capture and coding. As household response rates to surveys continue to fall, evaluations of the nonresponse bias in the survey estimates are becoming more common. These evaluations may involve postsurvey interviews of nonrespondents to determine why they were not interviewed and how their characteristics might differ from those of respondents. More often descriptive studies of the nonrespondents are conducted using whatever data are available on the nonrespondents, possibly from the sampling frame. In rare instances, embedded experiments and postsurvey evaluations might be conducted to estimate other components of the MSE such as reliability, measurement biases, or interviewer and other operator variances.

For surveys conducted on a continuing basis or repeated periodically, survey evaluation plays a critical role. It is important not only for continuous quality improvement but also as a vehicle for communicating the usefulness and limitations of the data series. For this reason, a number of US federal surveys have developed quality profiles (Biemer and Lyberg 2003) as a way of summarizing all that is known, as well as pointing out the gaps in knowledge, about the sampling and nonsampling errors in the surveys.

1.2.4 Latent Class Analysis

Latent class analysis (LCA), the main topic of this book, is a powerful method for estimating one or more parameters of a survey error model. These estimates can be used to evaluate the MSE of an estimate and the error probabilities associated with a survey question, as well as many other purposes. Application of latent class models is appropriate when the data to be analyzed are categorical (either nominal or ordinal categories). In that situation, the measurement errors in the observations are referred to collectively as *classification errors* or simply *misclassification*.

Evaluations of classification error are important because many survey variables are categorical. Demographic questions may ask individuals to classify themselves by gender, age, race, education, income level, tenure, marital status, and so on. Many survey questions are closed-ended—meaning that a small number of response options are presented from which the respondent must choose. The closed-ended response format is popular because it is simple for respondents to answer and responses are self-coded, making them easy to key as well. In addition, the information contained in the response options can indicate the type of responses the designer expects, which can often clarify the meaning of the question (Tourangeau et al. 2000). Responses to open-ended questions are seldom useful in data analysis unless they are coded to a reasonably small number of discrete categories prior to analysis. For example, a person's occupation is often collected as an open-ended response and then coded to one of a number of occupation categories. LCA is applicable in all of these situations.

As an example, consider the case where we want to classify respondents as either smokers or nonsmokers but our methods of questioning do not

accurately convey what we mean by smoking or some respondents may be unwilling to give us accurate information on their smoking habits. As a result, responses to a question on smoking habits are subject to misclassification. LCA is a statistical method for predicting the true classifications of individuals according to their observed classifications. In addition, LCA will provide estimates of the probabilities of respondents being misclassified by the question.

To apply LCA, we might ask two or three questions about smoking that are somewhat different but are all aimed at determining whether an individual smokes. LCA uses a statistical model that incorporates characteristics of the individual along with his or her responses to estimate a probability that the individual smokes. Under this model, an individual that answers "no" to all three questions still has a positive probability of being a smoker. LCA attempts to estimate the probability that individuals with varying patterns of responses to the smoking questions truly smoke. Each response pattern is associated with a probability of smoking that may vary depending on the other characteristics of the individual (age, race, sex, income, etc.). LCA will provide estimates of the true proportion of the population who smokes as well as error rates associated with each question about smoking.

However, as this book shows, many applications of LCA or its counterpart for panel surveys, Markov latent class analysis (MLCA), go beyond the estimation of classification error. Although this book focuses on measurement error, we also provide examples of applications of LCA for estimating census coverage error and for adjusting for nonresponse bias. New applications of LCA are still being discovered for this powerful statistical method.

1.3 ABOUT THIS BOOK

As implied by the title, the focus of this book is on the use of LCA as a tool for evaluating survey error. Survey error evaluations serve a number of purposes as noted in Section 1.2.1. LCA can be a valuable tool for all of these purposes. Although LCA finds application is evaluating all the error sources described in Section 1.1.3, our focus is primarily on measurement errors or, more broadly, content errors. A few applications are considered in Chapters 6 and 7 that aim to evaluate nonresponse and frame errors. These are included primarily to show the wide applicability of LCA. The book also focuses on error in cateogorical measurements (i.e., classification errors) rather than the errors in continuous measurements. The latter types of error are briefly treated in the next chapter to show the linkages with continuous and classification error models. This is not an important restriction because many of the variables used in survey data analysis are categorical. In fact, a common practice among survey analysts is to discretize continuous variables and treat them as either ordinal or nominal categorical variables in the analysis. This often leads to results that are more easily interpreted.

This book differs fundamentally from other books on LCA in that it focuses exclusively on survey error evaluation rather than more factor analytic purposes of LCA such as typological analysis, cluster analysis, and stage-sequential dynamics. Unlike more general uses of LCA, our framework assumes that the variable to be measured is well defined and is directly observable. The true value of the variable exists and can be measured accurately under ideal conditions. However, in the survey setting, it is measured with error. In this context, the primary objective of LCA is to estimate the magnitudes of the errors in measurements. These estimates can be used to identify problematic questionnaire items, to study the causes of the errors, and ultimately for survey improvement. A secondary objective is to estimate the prevalence of population characteristics that are corrected for classification error. The differences between LCA for survey error evaluation and general LCA methodology are further discussed and illustrated in Section 4.1.1.

The next three chapters trace the evolution of survey error evaluation from the early models of Hansen and his colleagues at the US Census Bureau to the modern and more sophisticated loglinear models with latent variables, of which the latent class model is a special case. In these chapters the linkages between the early models to the modern models is emphasized because a thorough understanding of LCA for survey error evaluation requires an understanding of its roots. It should be clear by Chapter 4 that LCA for survey evaluation is an extension of classical test theory concepts and ideas and the measurement error models of the 1950s and 1960s.

Chapter 2 describes a very general model for measurement error, one that has its roots in the early days of survey research. In that chapter, some of the basic concepts associated with the study of survey response and measurement error are described. A general model is described that can be applied to virtually any type of data. More useful models can be obtained by considering special cases of the model based on assumptions that are specific to either continuous or discrete data types. This chapter ends with a very simple model for binary data, where simple random sampling is assumed. The basic concepts form the foundations for the analysis of polytomous categorical variables under complex survey sample designs.

Chapter 3 picks up this thread and carries the ideas further into the realm of probability modeling when only two measurements of the true characteristic (i.e., the latent variable) are available. This situation is probably the most common in survey evaluations because obtaining more than two measurements of the same characteristic often requires additional efforts and resources. An interview–reinterview study, especially the *test–retest* reinterview (defined in Chapter 2), is one vehicle for obtaining remeasurements. Embedding remeasurements within the same questionnaire is also common; however, it must be used sparingly (say, for only a few characteristics) to avoid overburdening the respondent and unduly lengthening the interview. An advantage is that three or more replicate measurements of the same characteristics can be obtained in one interview. The basic ideas of LCA are also developed in this chapter.

Chapter 4 extends the latent class modeling concepts and specifications to $K \geq 2$ repeated measurements. As discussed in this chapter, an important advantage of adopting a latent class modeling framework for evaluating survey misclassification is the availability of general software for estimating the error components. Chapter 4 describes the standard latent class (LC) model and shows why its assumptions are unnecessarily restrictive for most applications. It introduces a more general modeling framework based on a loglinear model with latent variables and shows how the standard LC model is just a special case of the more general model. Within this general modeling framework, a wide range of error structures and error evaluation designs is discussed and analyzed using familiar loglinear modeling notation, methods and techniques. Chapter 5 considers some advanced topics in the study of LCA, including models for ordinal data and methods for complex surveys. Chapter 6 discusses some more advanced models and applications, including models for ordinal data and their use, the agreement model, and the capture–recapture latent class model.

Chapter 7 extends the ideas of LC models to panel survey data by considering Markov latent class (MLC) models. MLC models are important because they provide a means for estimating classification error directly from the data collected in the panel survey without special reinterview or response replication studies. Essential in the application of these models is some evaluation or assessment of the extent to which the model assumptions hold. These issues are discussed and the primary approaches for model validation are described.

Throughout these chapters, the emphasis is on the application of the models to survey error evaluations. In this regard, some topics that are relevant to general-purpose uses of latent class modeling (e.g., population typological analysis and latent transition analysis) are not treated in much detail. At the same time, topics often deemphasized in other discussions of LCA (such as error structure modeling) are emphasized in this book. Numerous examples and illustrations are presented to demonstrate the estimation methods as well as how to interpret the modeling results. The utility of the models for evaluating and improving survey data quality are discussed and demonstrated. The primary software package used for data analysis is the ℓEM software for fitting a wide range of LC models. This software is quite intuitive and can be downloaded from the Internet at no charge.

Finally, Chapter 8 summarizes the major innovations in the field of LCA and discusses the current state of the art. This discussion concludes by reviewing several issues where further research is needed and other matters that have been neglected in the previous chapters. Most of these issues are spawned from criticisms of the methodology over modeling assumptions and a general lack of rigor in prior applications of the methodology. These criticisms and how they can be dealt with in future applications are also discussed.