CHAPTER 1

Methods for Longitudinal Surveys

Peter Lynn
University of Essex, UK

1.1 INTRODUCTION

A *longitudinal survey* is one that collects data from the same sample elements on multiple occasions over time. Such surveys are carried out in a wide variety of contexts and for a wide variety of purposes, but in many situations they have considerable analytical advantages over one-time, or *cross-sectional*, surveys. These advantages have been increasingly recognised and appreciated in recent years, with the result that the number of longitudinal surveys carried out has multiplied. This growth of interest has occurred in the government, academic and private sectors.

For example, in the UK the Economic and Social Research Council (ESRC) has invested heavily in enhancing the country's already impressive portfolio of academic longitudinal surveys. The three existing long-term birth cohort studies that began life in 1946, 1958 and 1970 (Ferri *et al*., 2003) have been joined by a fourth, the *Millenium Cohort Study* (Dex and Joshi, 2005), which started collecting data in 2001 – and may soon be joined by a fifth. And an ambitious new household panel survey, *Understanding Society: the UK Household Longitudinal Study*, was commissioned in 2007 and begins data collection in January 2009 (www.understandingsociety.org.uk). With a target sample size of 40 000 households and a large boost of ethnic minority households this survey represents the largest single investment in academic social research resources ever launched in the UK. Meanwhile, the UK government has also initiated major new longitudinal surveys. The *Longitudinal Study of Young People in England (LSYPE)* started in 2004, involving annual interviews with a sample of over 15 000 young people born in 1989–1990 and their parents (http://lsype.notlong.com). The *Wealth and Assets Survey* entered the field in 2006 and plans to have biennial waves with a sample of over 30 000 persons. The *Offending, Crime and Justice Survey* (OCJS) began in 2003, with the main component being a longitudinal survey of young people aged 10–25, with annual interviews (Roe and Ashe, 2008).

Methodology of Longitudinal Surveys P. Lynn

This expansion in longitudinal surveys has also been visible in the International Arena. The *Survey of Health, Ageing and Retirement in Europe* (SHARE) began in 2004 with a first wave of data collection with a sample of over 30 000 individuals aged 50 or over in 11 countries (Börsch-Supan and Jürges, 2005). Further waves took place in 2006 and 2008. The *European Union Statistics on Income and Living Conditions* (EU-SILC) is a panel survey of households in all countries of the EU that began in 2004 and features a four-wave sample rotation pattern with annual interviews (http://eu-silc.notlong.com). The international suite of broadly comparable household panel surveys, which for years comprised only the UK, USA, Germany and Canada, has more recently been joined by Switzerland (1999), Australia (2001), New Zealand (2002) and South Africa (2008).

Reflecting this burgeoning interest in longitudinal surveys, and recognising that such surveys have distinct methodological features, this book aims to review the current state of the art in the design and implementation of longitudinal surveys. All aspects of the survey process are considered, from sample design through to data analysis and including overarching issues such as ethics and data linkage. This chapter aims to provide an overview of the features of longitudinal surveys that are distinct from other surveys and a nomenclature and framework within which to consider the other chapters.

1.2 TYPES OF LONGITUDINAL SURVEYS

The definition of a longitudinal survey presented in the first sentence of this chapter encompasses a multitude of different survey designs that are used in practice. Longitudinal surveys can vary on a number of key dimensions that tend to have implications for various aspects of methodology. These include features of the survey objectives, such as the study topics and the study population of interest (Binder, 1998), and features of the design such as the interval between waves, the mode of data collection, the treatment of new entrants to the population, and so on (Kalton and Citro, 1993). Some of the different types of longitudinal surveys are outlined below. This is not intended as a comprehensive typology, but merely to give a flavour of the variety that exists:

- *Surveys of businesses carried out by national or regional statistical offices*. These surveys tend to collect a limited range of information, typically restricted to key economic indicators. Data may be collected at frequent intervals, such as monthly or quarterly, and the main objectives are usually to publish regular series of statistics on totals, means and net change between periods, often for cross-cutting domains such as regions and industries. The rationale for a panel design is often to improve the precision of estimates of net change rather than an interest in estimates of gross change or any other micro-level longitudinal measures;
- *Surveys of school-leavers, graduates or trainees*. Institutions offering education or training, such as universities, or government agencies responsible for related policy, often wish to assess the outcomes of such education or training at a micro (student) level. These outcomes are often medium term or long term and consequently it is necessary to keep in touch with students/trainees for some time after they have completed their study. Longitudinal surveys are often used for this purpose, collecting

data from students/trainees on several occasions, perhaps beginning while they are still students/trainees and for up to several years after they have completed the course. The information collected is often quite complex, including perhaps full employment and activity histories between each survey wave and maybe also reasons for changes and decisions made. Sometimes, one or more 'control' groups may be included in the survey in an attempt to assess the impact of the education/training;

- *Household panel surveys*. In several countries, long-term panel surveys of the general household population are carried out. The oldest, the Panel Survey of Income Dynamics (PSID) in the USA, has been interviewing the same people since 1968 (Duncan *et al*., 2004). These surveys are multitopic and general purpose, collecting behavioural, attitudinal and circumstantial data on a range of social and economic issues. The main objective is to provide a rich data resource to be used by a wide range of users for a broad set of purposes. The data structure is complex, involving interviews with each person in the household of each sample member at each wave, in addition to household-level information and, often, additional survey instruments such as self-completion questionnaires or health measurements;
- *Birth cohort studies*. As already referred to in Section 1.1 above, the UK has a suite of four long-term birth cohort studies. Similar studies are carried out in many other countries. For some, the sample consists of all the births in a particular location (often a country) over a particular short time period such as a week. For others, such as the Millenium Cohort Study, a sample is taken of all births over a longer period, such as a year. Many of these studies began with a focus on either maternal health or child development (or both), but were later continued into the sample members' adult life when the value of doing so became apparent. Data about the sample members are typically collected from the mother in the first years of the study and later on from the sample members themselves;
- *Epidemiological studies*. Longitudinal medical studies often follow up samples of people over periods of time in order to study the progression of a disease or other changes in health, wellness and illness. The samples often consist of patients who have presented with a particular condition, either at a particular hospital or across a broader area such as the area of jurisdiction of a health authority. Some of these studies collect data through interviews or self-completion questionnaires while others collect only medical data via tests and diagnoses;
- *Repeated-measures studies*. This is a term found in the literature referring to studies that repeatedly measure the same variables in the same way on each sample unit. In practice, most longitudinal surveys of the types outlined above include some repeated measures, though they may vary in the proportion of the questionnaire or interview items that are repeated measures. For example, most items collected on household panel surveys are repeated (though not necessarily at every wave) in order to build up histories of economic activity, health status and so on (though an interesting development over the last decade or so is the recognition that simply repeating an item in identical fashion at every wave may not be the best way to obtain accurate measures of change or stability – see Chapters 5 and 6). Similarly, epidemiological studies take the same measurements repeatedly in order to chart disease progress. On the other hand, birth cohort studies often have very few repeated items, at least during the sample members' childhood years, as the questions that are appropriate to ask and the ways of testing development are somewhat age-specific. Surveys of school leavers or graduates tend to contain a mix of repeated items and wave-specific items;

- *Panels and cohorts*. The description above makes a distinction between household panel surveys and birth cohort studies. In fact the distinction is somewhat artificial and in practice consists only of differences of emphasis both in design and in substantive focus. A household panel contains samples of every birth cohort – but of course the sample of each cohort tends to be much smaller than would be found in a birth cohort study that concentrates on just a single cohort. And samples from a series of birth cohort studies can be combined to provide a picture of the situation of people of different ages at any one point in time. At the extreme, if a new birth cohort study were started every year, say, with a sample of births over a one-year period, then the combination of the studies would have the same population coverage as a household panel survey. The design focus of birth cohort studies, with a sample who are all very close to one another in age, tends to be associated with substantive research topics that are age-specific, such as growth and development. But there is no reason in principle why such topics could not be addressed with a household panel survey design, with age-specific survey instruments. The advantage would be that every age cohort is covered, providing some kind of population representation. The disadvantage is that in practice it would be impossible for any narrow age cohort to be afforded the kinds of sample sizes that are typically extant in birth cohort studies.

The connections between the two kinds of surveys may be even more obvious if we consider the case of surveys of graduates or school leavers. These typically involve samples from a one-year age group or stage group and therefore constitute another kind of cohort survey. Some of these surveys are carried out only once every n years, where n may take a value like 5 or 10. But some do indeed take place every year and thus, over time, they in principle cover the entire population of persons who have experienced the particular school system or educational establishment that is being sampled.

1.3 STRENGTHS OF LONGITUDINAL SURVEYS

Longitudinal surveys present a number of options, both for data collection and for analysis, that are either simply impossible with cross-sectional surveys or cannot be achieved in a satisfactorily accurate or reliable way. Often these features are key elements of the rationale for carrying out a longitudinal survey. However, it is important to distinguish between longitudinal *surveys* and longitudinal *data*. Longitudinal surveys are a source of longitudinal data, as the resultant data include items that refer to different points in time. But there are also other ways of obtaining longitudinal data, including diary methods and the use of retrospective recall within a single survey instrument. We briefly discuss here the strengths of longitudinal data, but our particular focus is on the strengths of longitudinal surveys.

1.3.1 Analysis Advantages

It is of course artificial to separate analysis advantages of longitudinal surveys from data collection advantages, as the reason for collecting a certain type of data is in order to be able to carry out certain types of analyses. The key advantages of longitudinal data

(which in most cases can only be accurately or adequately collected by longitudinal surveys) are analytical:

- *Analysis of gross change*. Analysis of gross change is perhaps one of the most common objectives of longitudinal surveys. Repeated cross-sectional surveys can be used to estimate net change, for example change in the proportion of employees who regularly cycle to work. But only a longitudinal survey can identify the extent to which this is composed of different elements of gross change. For example, suppose that the proportion of employees who regularly cycle to work is estimated to be the same at two points in time. It would be of interest to know whether it is exactly the same set of employees who cycle on both occasions or whether there are equal and opposite flows into and out of regular cycling (and, if so, how large they are and what the individual characteristics associated with each flow are, etc). These are the kinds of questions that longitudinal surveys can address.
- *Analysis of unit-level change*. Individual-level change can be of interest independently of interest in population-level net change. For example, analysts may wish to understand the characteristics and circumstances surrounding divorce regardless of whether there is net change in the proportion of people who are married.
- *Aggregate measures*. There may be a desire to derive measures that combine observations from multiple time points. An example might be combining 12 measurements of monthly expenditure to obtain an estimate of annual expenditure.
- *Measures of stability or instability*. Another reason for combining observations from multiple time points is to provide measures of stability or instability. Often individual-level change can only be well interpreted in the context of changes over a considerable period of time. With a multiwave panel that collects measures of income at each wave, it should be possible to identify people with different patterns of change, such as steady growth, fluctuation around a low level, sudden decline followed by stability, etc. The characteristics associated with such patterns are likely to be of considerable interest to policy makers. Patterns such as these can only be identified using accurate data referring to lengthy time periods. For example, poverty analysts have used household panel data to demonstrate that there is considerable instability over time in the poverty status of many individuals and households in Westernised countries (Jenkins and Rigg, 2001). While the proportion of households in poverty may remain relatively stable over time, there may be many entrants to and exits from poverty. A large proportion of households may experience at least one spell of poverty over a long period of time, while very few households may remain continuously in poverty throughout the period. This insight provided by longitudinal surveys may have shifted the policy focus from (stable) characteristics associated with poverty propensity at a point in time to better understanding poverty dynamics and the factors associated with falling into poverty or persistently failing to exit from poverty.
- *Time-related characteristics of events or circumstances*. Often characteristics such as frequency, timing or duration are of central interest to researchers. For example, understanding the duration of spells in a particular state and the factors associated with exiting from the state (sometimes referred to as 'persistence') are important for poverty, unemployment, marital and partnership status, participation in education and training, company profitability and many other topics. Hazard modelling and survival analysis are techniques used to better understand the propensity for change

(in any status of substantive interest) and the factors associated with such change. These techniques require longitudinal data.

- *Identifying causality*. Policy makers are ultimately interested in understanding what causes certain changes to occur. Most policies are designed to bring about change of some kind or other, but often informed by only limited knowledge of how a changed policy might have an impact on the desired outcomes. Thus, a key objective of researchers is to shed light on causality rather than mere association. Analysis of the ordinal nature of events, which requires longitudinal data, often sheds light on issues of causality. For example, a cross-sectional survey can establish an association between factors A and B. But a longitudinal survey might establish that, for most population units that have experienced both A and B, A happened before B, making it rather more likely that A caused B than *vice versa* (though of course a third factor, C, may have caused both A and B and this and other possibilities must always be considered). Chapter 17 discusses ways in which longitudinal surveys can help to understand causality.

A number of recent advances in methods for analysing longitudinal survey data are the topics of Chapters 16, 18, 19, 20, 21 and 22.

1.3.2 Data Collection Advantages

There are several ways in which longitudinal surveys provide advantages in terms of data collection. These are mostly connected with either the quantity or quality of data that can be collected compared with alternatives such as the use of retrospective recall. In many situations, the advantages are so clear that most researchers would conclude that a longitudinal survey is the only worthwhile option. In other situations, the advantages may be marginal. Additionally, there are certain specialised situations in which a longitudinal survey may have cost or logistical advantages over repeated cross-sectional surveys, even though the latter may in principle also be able to provide adequate data to address the research objectives.

Length of Histories

It is possible to collect much *longer* continuous histories of events and transitions with a longitudinal survey than could be collected retrospectively in a single interview, simply due to the volume of data involved (and hence the length of the interview or questionnaire).

Accuracy of Recall

Longitudinal surveys are often able to obtain more accurate data than could be collected in a single interview with retrospective recall if the data might be subject to recall error. This advantage accrues primarily because of the ability to limit the reference period over which respondents must recall information to a relatively short recent period – such as the past month, quarter or year – while ultimately obtaining data that relate to much longer periods. The reference period often equals the interval between waves, if the aim is to build up continuous histories. Recall accuracy tends to be associated with events and circumstances that are rare, salient and recent (Tourangeau *et al.*, 2000). For events and circumstances that do not have these characteristics, data recalled over long time periods

may be subject to severe recall bias. Recall error can take a number of forms. Events can be forgotten completely, they can be misplaced in time or misclassified, or associated characteristics may be misremembered. For example, most survey respondents are likely to be able to recall with reasonable accuracy the date(s) when they got married or had children, even if those events took place decades ago, but they are rather less likely to remember accurately when or how many times they visited a doctor, even if only asked about the past year.

Bounding

For many purposes, accurate dating of events is at least as important as accurate recall of the details of the event. But respondents may not be able to recall accurately the date of a specific event, even if they can recall the event itself. Consequently, retrospective recall questions asked in a single interview may produce biased estimates of frequencies and associated measures. A commonly reported phenomenon is 'telescoping', whereby survey respondents report events as having taken place within a reference period when in fact they took place longer ago. This is the result of errors in dating increasing the longer ago the recalled event took place (Rubin and Baddeley, 1989). Panel surveys have an extra advantage when collecting dates of events. Each interview after the first is *bounded* by the previous interview, so any events reported previously can be discounted from the reports in the current interview in order to avoid telescoping. This of course assumes that it can be unambiguously concluded whether or not reports in two consecutive interviews refer to the *same* event. Sometimes this is difficult, particularly when a respondent has a tendency to experience frequent events of a similar nature, but for many types of survey data it can usually be achieved.

Between-respondent Variation

The difficulty of any recall task will vary between respondents depending on the number, nature and timing of events they have experienced. This variation can be very considerable in the case of economic activity histories for example, causing a dilemma for survey designers. If a survey aims to collect complete activity histories over several years for a sample of people who are likely to vary greatly in their experiences, such as a cross-section of the general population, the ideal interval between survey waves will be very different for different sample members. But it is rarely possible to predict this in advance, nor is it often practical to have different between-wave intervals for different individuals. Instead, a standard interval is chosen. Interviews at annual intervals may be inefficient for persons whose circumstances change little (e.g. retired persons or those who remain in the same job for many years). The marginal amount of information collected in each interview, relative to the considerable cost, will be small. But annual interviews might present a considerable recall challenge to persons who experience large numbers of short spells of employment, perhaps interspersed with spells of unemployment or other activities. The advantage of longitudinal surveys in dealing with this challenge is that a lot of relevant information is known about sample members in advance of each wave (apart from wave 1). This information can be used to predict the likely complexity of recall tasks in various domains and tailored questioning strategies can be adopted. For example, those respondents who experience the highest frequency of changes in a particular domain may be administered a short telephone interview between waves in order to update information and thereby shorten the reference period. Or a group for

which particularly low levels of change are observed might be administered an extra module of questions on a different topic in order to maximise the value of the data collection exercise. Such tailored approaches can have disadvantages and will not always be advisable, but they are options that are not available to cross-sectional surveys. Thus, to gain maximum benefit from the ability of longitudinal surveys to collect longer and/or more accurate histories of events, the survey designer needs to understand the recall and reporting task being asked of respondents, how it relates to between-wave interval and how this relationship might vary over sample members.

Expectations and Intentions

Subjective measures such as expectations, intentions, attitudes and reasons for making particular choices may be particularly prone to recall error. Though some researchers have retrospectively collected attitudes (Jaspers *et al.*, 2008) or reasons for making decisions (De Graaf and Kalmijn, 2006), others believe that attempting such measurement is simply not worthwhile, in particular because respondents' memory of such things is likely to be tainted by subsequent experiences and post hoc rationalisation may lead respondents to reassess their earlier motivations or expectations (e.g. Smith, 1984). Longitudinal surveys are able to collect information about expectations and reasons for choices untainted by subsequent events and outcomes. Later waves can of course also collect information about the relevant subsequent events and outcomes. This enables analysis of associations between, say, expectations and outcomes, that would simply not be possible with any other data source.

Maintenance of Samples of Rare Populations

For certain rare study populations, the cost of assembling a representative sample may be very high relative to the cost of collecting survey data from the assembled sample. This may be the case for populations that are not identified on any available sampling frame and for which it is therefore necessary to carry out a large-scale population screening exercise (Kalton and Anderson, 1986). In such cases, it may be cost-efficient to carry out repeated surveys with the same sample rather than select a new sample for each new study of the same population. The motivation in this case, therefore, is not a need for longitudinal data but the result is nevertheless the use of longitudinal data collection methods.

1.4 WEAKNESSES OF LONGITUDINAL SURVEYS

1.4.1 Analysis Disadvantages

Longitudinal surveys are often not as good as cross-sectional surveys at providing cross-sectional estimates. This may be perceived as a weakness, but it is simply not something that longitudinal surveys are designed for. Compared with estimates from a cross-sectional survey, cross-sectional estimates from a longitudinal survey (from wave 2 onwards) may be more likely to suffer from coverage error if the sample does not include, or under-represents, additions to the population since the sample was selected. This coverage error may increase over waves as the time since sample selection, and hence the undercoverage rate, increases. The design of a longitudinal survey can often be adjusted

to improve the quality of cross-sectional estimates that can be made, though this may be resource-intensive and may detract from the central longitudinal aims of the survey. Such design options are discussed in Chapter 2. Also, a longitudinal survey sample may suffer from a lower net response rate than a cross-sectional survey on top of the lower coverage rate (though lower coverage and response rates do not necessarily imply greater error: this depends on the association between the coverage and response propensities and key survey estimates).

A difficulty with longitudinal analysis may be the analysis-specific nature of the population to which estimates refer. This can create problems in understanding and communicating effectively the nature of the reference population and consequent difficulties of inference. Definitions of study populations are discussed further in Section 1.5 below.

1.4.2 Data Collection Disadvantages

There are some aspects of survey data collection that are unique to longitudinal surveys and potentially detrimental or problematic.

Panel Conditioning

Panel conditioning refers to the possibility that survey responses given by a person who has already taken part in the survey previously may differ from the responses that the person would have given if they were taking part for the first time. In other words, the response may be *conditioned* by the previous experience of taking part in the survey. This therefore relates to all data collected by longitudinal surveys other than those collected at the first wave. There are two ways in which conditioning can take place. The way in which respondents *report* events, behaviour or characteristics might change; or the actual *behaviour* might change.

For example, a two-wave survey of unemployed persons might find that more people report a particular type of job search activity at the second wave than at the first wave. This might reflect a genuine increase in the extent to which that activity takes place (independent of taking part in the survey) but it could also be affected by panel conditioning. This could be because the first wave interview made some sample members aware of possible job search activities in which they were not currently engaged, so they subsequently started doing those things. Thus, there was a genuine increase in the extent of the activity, but only amongst sample members – not amongst the population as a whole. The *behaviour* of the sample members has been conditioned by the experience of the first interview. Alternatively, the experience of the first interviews may have affected the way in which some sample members respond to the questions in the second interview, even though their actual job search behaviour may not have changed. Perhaps in the first interview they discovered that reporting no activity of a particular type led to them being asked a series of questions about why they did not participate in that activity. So, to make the second interview shorter, or to avoid embarassing questions, they now report that they have participated in this particular activity. In this case, the *reporting* of the sample members has been conditioned by the experience of the first interview.

In the above example the conditioning was caused by having been asked the same question previously, but in other cases the conditioning could be caused merely by having been interviewed previously. Questions that were not asked previously can therefore also

be affected. One process by which this might happen is that respondents could build up a greater sense of trust of the interviewer and the survey organisation as time passes. If asked some questions on a sensitive topic at wave 3, say, they might then be more willing to reveal truthful answers than they would have been at wave 1. This example also illustrates that some forms of conditioning can result in improvements to data quality. It is not the case that conditioning always damages quality. Another way in which quality might improve over waves of a panel is if respondents learn that it is useful to prepare for the interview by assembling certain documents. Examples of such documents might be pay slips, utility bills, health records or examination results, depending on the survey topics.

Panel conditioning is more likely to occur for certain topics and certain types of questions than for others, meaning that the extent to which the researcher needs to be concerned about panel conditioning – and the measures that might be introduced to minimise any negative consequences of it – should depend on the survey content. Panel conditioning in the context of attitudinal measures is discussed in Chapter 7.

Sample Attrition

Sample attrition (also referred to as *panel attrition*) refers to the continued loss of respondents from the sample due to nonresponse at each wave of a longitudinal survey. The term attrition is usually used to refer to a monotone process whereby sample members can change from being respondents to nonrespondents but not vice versa. In fact, many longitudinal surveys continue to attempt to collect data from sample members after they have been nonrespondents at one or more waves, and are often successful in doing so. Generally, then, the issue is that many sample members may have been nonrespondents at one or more wave, whether or not their response pattern is one of monotone attrition. The response rate at any one wave of a longitudinal survey may be just as good as that for any other survey but after, say, five waves the proportion of sample units that have responded at every wave may be quite low. Thus, the effective response rate for longitudinal analysis – for which data from every wave is required – may be lower than the response rates that we are used to having on cross-sectional surveys. After several waves there is also a risk that responding sample sizes can become unacceptably small. Minimising sample attrition is consequently a major preoccupation of researchers responsible for longitudinal surveys. The factors affecting sample attrition are discussed in Chapter 10. Chapters 11, 12 and 13 discuss various strategies for minimising attrition while Chapters 14, 15 and 16 address issues of how to deal appropriately with attrition at the analysis stage.

Initial and Ongoing Costs

The time dimension of a longitudinal survey brings with it a set of extra tasks for the survey organisation. In particular, the design and planning work that must take place prior to wave 1 is considerably greater than that required prior to a cross-sectional survey. The wave 1 instrument(s) cannot be designed independently of the instruments for later waves as it is the combination of data from each wave that will provide the necessary survey estimates. Sample design and fieldwork planning must take into account the need to follow sample members throughout the life of the survey. An integrated survey administration system is needed so that both field outcomes and survey data from each wave can determine the actions to be taken at later dates (these may include things like keeping-in-touch exercises – see Chapter 11 – and dependent

interviewing – see Chapters 5 and 6). The sample management system needs to be able to cope with updated contact information arriving at any time and needs to be able to log the history of changes in such information. If the survey involves face-to-face interviewing, it becomes impossible with a longitudinal survey to maintain the cost advantages of a clustered sample, as the sample will disperse to some extent over time even if it is clustered initially. These considerations tend to result in the costs of an n-wave longitudinal survey exceeding the costs of n cross-sectional surveys, holding sample size and mode constant, though this need not always be the case.

1.5 DESIGN FEATURES SPECIFIC TO LONGITUDINAL SURVEYS

There are some aspects of survey design that are unique to longitudinal surveys, or are substantially different in nature to cross-sectional surveys. Standard survey methods textbooks provide little or no guidance on how to make design decisions regarding these aspects. Yet these aspects warrant careful consideration as design decisions can have weighty implications for data quality and for analysis possibilities.

1.5.1 Population, Sampling and Weighting

Population Definition

As data collection takes place at multiple time points for each sample unit, the data themselves are defined by time. Consequently, the time dimension must also enter the definition of the population being studied. To estimate a quantity such as mean change between time point 1 (t_1) and time point 2 (t_2), say $\hat{\overline{\Delta}}_i$, where $\Delta_i = Y_{it_2} - Y_{it_1}$, the analyst requires a sample from the population of units that existed at both t_1 and t_2, say $P_{t_1} \cap P_{t_2}$. Quantity Δ_i is not defined for any units that existed at only one of the two time points. In consequence, for any longitudinal survey with more than two waves it will almost certainly be the case that the study population will depend on which waves provide data for an estimate. Therefore different subsets of the sample units will represent different populations of interest and will thus be relevant for different estimation purposes. For example, suppose the time points refer to survey waves. An estimate that requires data from waves 1, 2 and 3 can only refer to the population $P_{t_1} \cap P_{t_2} \cap P_{t_3}$, which is a subset of the population $P_{t_1} \cap P_{t_2}$ to which our estimate of $\overline{\Delta}_i$ refers. Clarity is needed about the populations to which inference is required. This has implications for sample design (see Chapter 2), for following rules and for weighting (see below, and Chapter 15 for a discussion in the context of household panel surveys).

Sample Design

Longitudinal surveys face a number of unique sample design issues. Decisions must be made as to whether and how new entrants to the study population (births) should be included, whether a rotating design should be used, whether a repeated design should be used and, if so, whether the repeats should overlap or whether a split design is desirable. These fundamental design options are discussed in Chapter 2.

Additionally, some other features of sample design take on a rather different complex in the context of longitudinal surveys. If face-to-face interviewing is to be used at any or

all waves, geographical clustering of the sample might be desirable on cost efficiency grounds. However, the cost parameters are more complex and harder to estimate as they depend on the extent and nature of sample mobility between face-to-face waves. Predictions of design effects should also influence decisions about sample clustering, but there is much less knowledge about design effects for clustered longitudinal surveys – especially for longitudinal estimates – than there is for cross-sectional surveys. To the extent that intracluster correlations are caused by shared personal characteristics, these should persist even as sample clusters disperse geographically over time. But to the extent that geography per se influences survey measures, dispersal might tend to reduce intracluster correlations. This issue warrants further investigation.

Surveys often oversample population subgroups that are of particular interest. For a cross-sectional survey, this is a useful device and is unproblematic provided that relative selection probabilities are known. For a longitudinal survey, the considerations are a little more complex. If subgroup membership is defined by characteristics that are variable over time, such as income level or geographical location, then oversampling the subgroup at wave 1 can become an inefficient strategy at later waves when substantial numbers may have moved into and out of the subgroup. A particular case of this occurs on business surveys, where businesses may, during the course of their inclusion in a panel survey sample, move between strata that were sampled at very different rates (for example, if the strata are defined by number of employees). Even if subgroup membership is defined by time-invariant characteristics, there is a risk that the focus of interest could shift over the duration of the survey, such that the sample becomes inefficient for current research interests.

Weighting

Many longitudinal surveys provide data users with two types of weights – *cross-sectional weights* and *longitudinal weights*. The distinction reflects a difference in the population to be represented, which in turn is related to different estimation objectives, and in the sample cases that can contribute to estimates. Given the focus of longitudinal surveys on longitudinal estimation, the longitudinal weights are perhaps the more important of the two. However, it is not always obvious for which sets of respondents these should be produced.

There are $2^t - 1$ possible populations that can be represented by a t-wave longitudinal survey, of which t are cross-sectional populations and $2^t - (t + 1)$ are longitudinal populations. Potentially, a set of weights could be created for each population. However, for surveys with more than two or three waves it would not be feasible to create all these sets of weights. It may be confusing to users to have so many sets of weights available and it is probably not necessary anyway, as many of the sets of weights would be so similar that the choice between them would make no practical difference to any estimates.

A solution would be to provide users with the data necessary to calculate weights for any combination of waves – and some guidance or even a program that will calculate the weights. Then, each user could produce weights tailor-made to his or her analysis. However, this is rarely done, either because some of the necessary data cannot be released at the unit level or because users who want to use weights prefer to be provided with ready-to-use weights.

An alternative is for the data provider to produce weights for a limited subset of the possible combinations of waves. This should be accompanied by guidance to users on

what to do if the combination in which they are interested is not one of those for which weights are provided. The choice of wave combinations should be guided by the (likely) main uses of the data. For example, if the main objective of the survey is to permit analysis of change relative to baseline data that were collected at wave 1, then there is very little point in producing weights for combinations of waves that do not include wave 1. If a module of questions on a particular topic is included only at waves 1, 4, 7 and 10, then that particular combination should be a strong candidate for weighting. For almost all longitudinal surveys, the complete set of waves should be one of the combinations for which weights are produced. The only exception would be if, by design, there are no units eligible for data collection at every wave.

The choice of sets of weights to produce can have a real impact, as analysis based on respondents to a particular set of waves using weights designed for a different set of waves is suboptimal. Consider a three-wave survey for which one set of longitudinal weights is provided, designed to make the set of persons who responded to all three waves representative of $P_{t_1} \cap P_{t_2} \cap P_{t_3}$. Suppose we want to estimate a parameter of change between wave 1 and wave 3, for which we only need to use data collected at waves 1 and 3. We could use sample members who responded only at waves 1 and 3, in addition to those who responded at all three waves. However, the longitudinal weights will be set to zero for the former group, so these cases will be dropped from the analysis.

Another important consideration is that weights are typically produced after each new wave of data becomes available. Thus, as a minimum, at each wave a set of weights will be produced representing the longitudinal population at all waves to date. This means that ultimately weights will be available for every *attrition sample*. If the survey policy is to attempt to collect data only from previous wave respondents this will be all the weights that are needed. Otherwise, the task is to identify which other combinations of waves are sufficiently important to warrant the calculation of weights.

1.5.2 Other Design Issues

Intervals between Waves

The frequency of data collection waves has several implications. For recall data, higher frequency waves are likely to produce higher quality data, as discussed in Section 1.3.2 above. However, the optimal frequency will depend on the nature of the information to be recalled. For indicators of current circumstances, where the intention is to identify changes in circumstances, the optimum frequency will depend on the rate of true change. If waves are too infrequent then important changes might be missed, while if they are too frequent then expensive data collection efforts will yield little useful information.

Nonresponse should also be a consideration in the choice of between-wave interval. For a given overall survey duration (see below), a shorter between-wave interval (and therefore more waves) is likely to lead to higher nonresponse, in the sense that a higher proportion of sample members are likely to be nonrespondents at one or more wave. But for a given number of waves, a shorter interval is likely to be associated with lower nonresponse, not least because smaller proportions of sample members will move home or otherwise change their contact details, rendering it easier to locate and contact them over multiple waves. For a given overall survey duration, a shorter interval also has obvious cost implications as there is a cost associated with carrying out each extra wave.

Duration of the Survey

The longer the period of time over which a longitudinal survey collects data, the richer and more valuable the data are likely to be. However, for some surveys the focus is limited to a particular phase in the life of sample members, for example defined by age, education or employment, so this might naturally imply a limited duration for the survey. A longer duration will have extra costs, independent of the number of waves carried out during that time. These arise due to the need to maintain staffing and to perform tasks such as panel maintenance and tracking, which are independent of data collection tasks.

Respondents and Study Units

In most of the examples mentioned in Section 1.1 above, data are collected from, and about, individuals. In some cases, and for some analysis purposes, the units of interest may be groupings of individuals such as couples, households or parent–child groups, but even then many of the group-level variables will be constructed from individual-level data items, though some may have been collected from a single individual in the group but referring to the group as a whole. Indeed, on all surveys the respondents are individuals, but in some cases they are supplying data that refer to some other entity, such as a business, farm, educational establishment, community or other entity. Longitudinal surveys where the study unit is an establishment of some kind are common (see Cox *et al.*, 1995; American Statistical Association, 2000). The relationship between respondent and study entity has important implications for such surveys: there may be multiple respondents at each wave; the survey organisation may have to take extra measures if they want to know who the respondents were; and respondents might change from wave to wave, for example through staff turnover. These respondent issues may have implications for measurement error as well as implications for contact, motivation and cooperation. On establishment surveys, the changing nature of establishments over time may also create issues for population and sample definition. Establishments can change name, size, nature and even location. A longitudinal survey needs clear definitions of what constitutes a continuing unit over time as opposed to a unit that has ceased to be part of the study population (a *death*) or one that has entered the population (a *birth*).

Tracking and Tracing

The long-term health of a longitudinal survey is dependent on retaining the ability to contact sample members at each wave. This requires an administrative system and a programme of operations to be put in place. Such operations might include regular mailings or e-mails, the collection and maintenance of contact details of close friends and relations of sample members, and linking sample details to administrative data sources that indicate address changes. Some of the challenges in tracking sample members and the outcomes of some commonly used techniques are discussed in Chapter 11.

Modes

There exists a wide range of methods of collecting survey data. Respondents may be interviewed either face to face or by telephone, or self-completion questionnaires may be administered. In each of these cases the instrument itself may or may not be computerised. In the face-to-face case, visual prompts such as show-cards may or may not be

used. In the self-completion case, options exist for the method by which the questionnaire is delivered to, and collected from, respondents. All of these options have implications for costs, timetables, response rates and for data quality and survey error. The choice of data collection modes is therefore important. In the case of longitudinal surveys, this choice has some extra dimensions.

First, a greater variety of ways of combining modes is available. Sample members can be switched between modes from one wave to the next, in addition to any combination that may be done within a wave (for example, a self-completion component within a face-to-face interview). A particularly important consequence of this is that it may affect measures of change between waves if the modes in question have different measurement properties. This issue is the focus of Chapter 8.

Second, information can be collected at one wave that opens up extra mode possibilities at future waves. For example, respondents can be asked to provide e-mail addresses, allowing the possibility of an approach by e-mail to take part in a web survey at the next wave. However, these opportunities tend to lead to mixed-mode data collection (within a wave) as the proportion of sample members supplying a valid e-mail address or phone number, etc. rarely approaches 100 %.

1.6 QUALITY IN LONGITUDINAL SURVEYS

The ideas of survey quality (Biemer and Lyberg, 2003) provide a set of tools for maximising the usefulness of any survey. Within this, the concept of total survey error (Groves, 1989) can provide a useful framework for considering the statistical impacts of design and implementation considerations for longitudinal surveys. The total error in any survey-based estimate, $\hat{Y}$, of a population parameter, Y, is simply $\hat{Y} - Y$, the difference between the estimate and the true value of the parameter. Many factors can contribute to this error. The main components of error in any survey estimate are set out in Figure 1.1. We outline in turn key influences on each in the longitudinal survey context.

1.6.1 Coverage Error

Aside from undercoverage in the initial sampling frame, which is an issue common to all surveys, the main source of undercoverage on longitudinal surveys may be a failure to (adequately) include new population entrants in the sample. The extent to which this is a problem will depend on the definition of the study population as well as the survey sampling procedures.

Sources of Statistical Error in any Survey Estimate	
Errors of Nonobservation	**Observational Errors**
Coverage	Measurement
Sampling	
Nonresponse	

Figure 1.1 Components of total survey error.

1.6.2 Sampling Error

The design issues that affect sampling error are not distinct for a longitudinal survey, other than the considerations regarding clustering and regarding differential selection probabilities, discussed above in Section 1.5.1. It should always be remembered, however, that the nature of the estimates for which sampling error is of interest is typically rather different from the nature of cross-sectional descriptive estimates. Knowledge of sampling error in one context cannot therefore be easily translated to the other context.

1.6.3 Nonresponse Error

In addition to the processes that can lead to nonresponse on any survey (Groves and Couper, 1998), there are some important influences on response propensity that are rather distinct in the longitudinal survey context.

Subsequent to the first wave, a major component of *noncontact* on longitudinal surveys is caused by geographical mobility of sample members. Between waves, some sample members will move home, change employment, change telephone number or change e-mail address. If the survey organisation is reliant upon any or all of these details to contact the sample member, then they will need to take extra measures to be able to make contact at the subsequent wave. Techniques for tracking sample members between waves are discussed in Chapter 11. For a face-to-face interview survey it is also necessary to have interviewers available to attempt contact with sample members in all the places to which they could move.

Participation in a longitudinal survey requires considerable commitment from sample members – not just a single interview, but several, over a period of time. In consequence, special incentives or motivation may be needed to compensate. Typically, sample members are offered a small payment for each interview, or some other form of gift, and particular effort is made to make the sample member feel an important, irreplaceable, component of the study and to persuade them of the value of the study. The use of incentives is discussed in Chapter 12.

A unique feature of longitudinal surveys relevant to *refusals* is that, after the first wave, sample members have already experienced the survey interview and therefore have a very good idea of exactly what it consists of, what kinds of questions will be asked, how difficult or sensitive they find it, and so on. This is very different from a typical survey situation, where a sample member will have only a rather vague and general impression of what they are being asked to do at the time when they are being asked to cooperate. Consequently, on a longitudinal survey it is very important to try to make the interview experience as pleasant as possible for the respondent. If a respondent finds the interview difficult, frustrating, embarassing, threatening, uninteresting or simply too long, they will be less likely to be willing to take part again at the next wave. Chapter 10 discusses sample attrition.

1.6.4 Measurement Error

Measurement error refers to the possibility that any individual survey observation might differ from the value that would be observed by a perfect measurement of the concept that the survey item was intended to measure. Such errors can arise for many reasons, associated with the response task, the respondent, the interviewer, the data collection mode and the interview setting (Biemer and Lyberg, 2003, chapter 4), but some features of measurement error are distinctive in the case of longitudinal surveys.

Of central interest to longitudinal surveys is the estimation of micro-level change over time. As measures of change draw upon data collected at different waves, any differences in measurement properties between waves or inconsistencies in the observed measures between waves can cause particular problems. Recall error in retrospective data can contribute to a phenomenon that is only observable in data from longitudinal surveys, known as *seam bias* or *seam effects* (Jäckle and Lynn, 2007; Jäckle, 2008). This refers to a common finding that the level of observed change between two consecutive time periods is much higher when the observations for each period come from two different interviews than when they come from the same interview. In other words, there appears to be an excess of transitions at the 'seam' between reference periods. This phenomenon can obviously introduce error to estimates of levels of change, but also to estimates of the correlates or causes of change, etc. Methods that reduce the extent of seam effects, and hence reduce error in estimates of change, are therefore desirable. Identification of such methods is, however, not entirely straightforward and requires understanding of the range of manifestations of recall error and their causes (Jäckle, 2008). *Dependent interviewing* encompasses a range of questioning techniques that take advantage of the bounding possibilities of longitudinal surveys described in Section 1.3.2 above and is discussed in Chapters 5 and 6.

When the study of change is based not on recall data but simply on comparisons of 'current status' indicators collected at each wave of a longitudinal survey, measurement error can still cause considerable problems. Even random misclassifications, for example, can bias estimates of change. Some analytical solutions to this problem are offered in Chapter 21.

Many aspects of survey methods can affect the measurement properties of a survey item, including whether the question and response options are presented visually or aurally, whether an interviewer is present, the setting of the interview, the context of the item, and so on. Any change between waves in survey methods therefore has potential to alter the measurement properties and hence to introduce error to estimates of change. Researchers should be particularly wary of design changes that have been shown to influence measurement for relevant topics or question types. Examples might include the effect of interviewer presence for items that have a social desirability connotation (Aquilino, 1994; Hochstim, 1967) or for which respondents may show a tendency to satisfice, the effect of aural vs. visual presentation for cognitively complex response construction tasks (Krosnick, 1999), and the effect of question context for attitude questions (Schuman and Presser, 1981; Tourangeau and Rasinski, 1988). Some of these measurement issues are explored in Chapter 8.

Even if survey methods are held completely constant between waves, there is still a risk that the measurement properties of an item for a given respondent can change between waves due to panel conditioning, as introduced in Section 1.4.2 above (and see Chapter 7). Methods should therefore be sought either to minimise the negative aspects of panel conditioning or to control for it in analysis. The latter approach is likely to require information from either a split panel or rotating panel design (see Chapter 2).

1.7 CONCLUSIONS

Longitudinal surveys are subject to all the methodological considerations that apply to any survey (Groves *et al.*, 2004) but they also have an added dimension that brings a wealth of extra opportunities but also some extra complexities and problems. This

volume aims to provide an overview of those issues with the hope of promoting best practice in the design and administration of longitudinal surveys and encouraging clear thinking regarding the many issues that a longitudinal survey researcher must deal with. Close reading of many of the chapters will, however, reveal many areas in which the research evidence to inform good design decisions is rather thin. A secondary hope is therefore that this book will stimulate researchers to further investigate some of the many fascinating methodological issues surrounding longitudinal surveys.

REFERENCES

American Statistical Association (2000). *ICES: Proceedings of the Second International Conference on Establishment Surveys: June 17–21, 2000 Buffalo*, New York. Alexandria, VA: American Statistical Association.

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: a field experiment. *Public Opinion Quarterly*, 58, 210–240.

Biemer, P. P. and Lyberg, L. E. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons, Inc.

Binder, D. (1998). Longitudinal Surveys: why are these surveys different from all other surveys? *Survey Methodology*, 24, 101–108.

Börsch-Supan, A. and Jürges, H. (2005). *The Survey of Health, Ageing and Retirement in Europe – Methodology*. Mannheim: MEA. Retrieved October 7, 2008, from http://share-methods.notlong.com

Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J. and Kott, P. S. (1995). *Business Survey Methods*. New York: John Wiley & Sons, Inc.

De Graaf, P. M. and Kalmijn, M. (2006). Divorce motives in a period of rising divorce: evidence from a Dutch life-history survey. *Journal of Family Issues*, 27, 483–505.

Dex, S. and Joshi, H. (2005). *Children of the 21st Century: From Birth to Nine Months*. London: Policy Press.

Duncan, G., Hofferth, S. L. and Stafford, F. P. (2004). Evolution and change in family income, wealth, and health: The Panel Study of Income Dynamics, 1968–2000 and beyond. In J. House, *et al*. (Eds), *A Telescope On Society* (Chapter 6). Ann Arbor: Institute for Social Research, University of Michigan.

Ferri, E., Bynner, J. and Wadsworth, M. (2003). *Changing Britain, Changing Lives: Three Generations at the Turn of the Century*. London: Institute of Education.

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

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

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. New York: John Wiley & Sons, Inc.

Hochstim, J. R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976–989.

Jäckle, A. (2008). The causes of seam effects in panel surveys. *ISER Working Paper* 2008–14. Colchester: University of Essex. Retrieved October 13, 2008, from www.iser.essex.ac.uk/pubs/workpaps/pdf/2008–14.pdf

Jäckle, A. and Lynn, P. (2007). Dependent interviewing and seam effects in work history data, *Journal of Official Statistics*, 23, 529–551.

Jaspers, E., Lubbers, M. and De Graaf, N. D. (2008). Measuring once twice: an evaluation of recalling attitudes in survey research. *European Sociological Review*. Advance access published online 13 October 2008.

Jenkins, S. P. and Rigg, J. A. (2001). The dynamics of poverty in Britain. *DWP Research Report*, No. 157. London: Department for Work and Pensions.

Kalton, G. and Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society Series A*, 149, 65–82.

Kalton, G. and Citro, C. F. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology* 19, 205–215.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537–567.

Roe, S. and Ashe, J. (2008). Young people and crime: findings from the 2006 Offending, Crime and Justice Survey. *Home Office Statistical Bulletin* 09/08. London: Home Office. Retrieved October 7, 2008, from http://www.homeoffice.gov.uk/rds/pdfs08/hosb0908.pdf

Rubin, D. and Baddeley, A. (1989). Telescoping is not time compression: a model of the dating of autobiographical events. *Memory and Cognition*, 17, 653–661.

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

Smith, T. (1984). Recalling attitudes: an analysis of retrospective questions on the 1982 General Social Survey. *Public Opinion Quarterly*, 48, 639–649.

Tourangeau, R. and Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299–314.

Tourangeau, R., Rips, R. and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.