4

Doing secondary analysis

> Perhaps the most important attribute for the user of published data is a large dose of scepticism. (Jacob 1984: 45)

This chapter considers some of the things a researcher needs to think about when using secondary data. There is a misapprehension among some commentators that doing secondary analysis is easy because it bypasses the tricky data collection phase (Glaser 1962). Neither is it necessarily the case that data researchers did not collect themselves are more prone to error than primary data. Rather, secondary data suffer from the same limitations that affect all social data and researchers need to make the same analytic decisions and compromises they would make when preparing and analysing data they collected themselves. It is, of course, true that there are specific things that the analyst needs to think about when using secondary data and it is these that we focus on in this chapter. We begin by considering the need to establish the purpose of the secondary data source being considered as well as its relevance to one's research questions. The discussion then considers practical issues such as the resources need to undertake the analysis and technical issues such as sampling, response rates and the nature of the variables included in the analysis and, in particular, their fit with one's own research questions. These considerations are then brought together in a worked example that explores the potential of the General Household Survey for use in a secondary analysis of household consumption.

What is the purpose of the research?

> The successful secondary analyst must be able to use and interpret the data with the knowledge and insight that went into its original collection. (Dale et al. 1988: 16)

With this in mind, a good place to start is, rather obviously perhaps, to find out exactly what it was that the primary research was trying to achieve. You might begin by examining the research questions or aims of the research or look at who commissioned and undertook the study: was the study commissioned by a government department, was it study based on opinion polls or market research or was it a piece of academic research? The easiest way to find this out is to examine the research manuals or other documentation accompanying the study or, if the data are presented as a report, this information should be located in accompanying appendices.

Because secondary data are not necessarily more objective than data based on observations or ethnographies, it is important to get some idea about the concepts that motivated the original research and which may influence its application to your secondary analysis. If the research was commissioned or undertaken by an advocacy group, for example, there may be a particular aspect or interest that the research emphasises. This narrower focus may be fine for advocacy research but might be insufficiently objective to be presented as evidence in an academic study. For example, polls on abortion law reform commissioned by right to life campaigners and pro-choice groups that have been conducted by the same polling companies over similar time periods are still able to produce very different results – results that may reflect the ideology of the particular commissioning group (Finney and Peach 2004: 14).

Another thing to think about when selecting your data is the type of conceptual or theoretical framework that was used in the original study. Often this is apparent from reading the manuals or appendices which accompany the data. Again, caution is needed as apparently very similar surveys can have subtle but important differences. For example, two widely cited international studies of student attainment the *Trends in International Mathematics and Science Study (TIMSS)* and the *Programme for International Student Assessment (PISA)*, have very different conceptual frameworks. PISA is concerned with understanding young people's ability 'to use their knowledge and skills in order to meet real-life challenges rather than how well they had mastered a specific school curriculum' (PISA 2000: 16), while, in TIMSS, the focus is on the curriculum, on what the curriculum contains, how it is implemented and what it achieves or, in other words, what students have actually learnt (Mullis et al. 2004). Although these differences are very subtle – they are

still measuring knowledge and understanding of the school curriculum – the secondary analyst will need to decide which study is closer to their own theoretical or substantive interests.

Once the purpose of the study is apparent, there are a few other preliminary checks that the analyst might wish to make on their chosen study: they concern the method of data collection and the structure of the questions used.

Who collected the data?

Often the data produced by large-scale surveys is collected through face-to-face interviews between the respondent and a professional interviewer who may have no other relationship with the study. Therefore, it is worth finding out who collected the data. If professional interviewers were used, they are likely to have followed a scripted questionnaire. The guidelines that are given to interviewers conducting such surveys can be very precise. For example, in the detailed script for interviewers conducting the 2002 World Values Survey, interviewers were required to show cards at appropriate points, reverse the order of options for question responses and not to read out the 'don't know' options (ICPSR 2005). Contrast this perhaps with academic researchers who might unthinkingly rephrase a question to ensure it has been properly understood (Porter 1995) and so may elicit a different type of response or develop a different relationship with the respondent than an interview that kept to a strict script. Neither way of collecting data is necessarily problematic for the secondary analyst, however, it is still important to know who undertook the survey in order to help gauge how objective the research might be.

How were the data collected?

It is also important to find out how the data were collected. Many of the large-scale surveys that are described in Appendix 1 adopt traditional data collection techniques, such as questionnaire and/or face-to-face interview, but the data collected by market research or opinion poll companies may have employed different methods. Data generated by these companies are not necessarily less methodologically rigorous than academic surveys (Harrop 1980), but they may lack substantive rigour: some polls have an astonishingly quick turnaround, often as short as 24 hours, leaving little time for question development and data analysis (Finney and Peach 2004). In the UK the recent success of online polling companies such as *YouGov*, further underlines the importance of knowing where your data came from. Although claiming validation for their method by recent successes in predicting the outcomes of events as

diverse as the 2003 elections to the Scottish Parliament and the first Pop Idol contest (Kellner 2004), 'internet polling is still susceptible to a number of biases and methodological weaknesses' (Finney and Peach 2004: 16). Secondary users of such data will need to be alert to issues such as low response rates, the use of incentives to participate, as well as issues of sampling bias, in particular towards those who have access to the internet and who would choose to register and volunteer to participate in these surveys. In addition, because we know very little about whom in the general population has access to the internet, we cannot draw a probability sample and this limits the generalisability of any findings (Sparrow and Curtice 2004). By the same token, there are also flaws with traditional polling methods: telephone and face-to-face interviews are biased towards those who might be at home at the right time; databases of landline telephone numbers exclude mobile numbers and answer phones and voicemail services enable people to screen calls and choose who they wish to talk to (Kellner 2004). The message for the secondary analyst is that all methods of data collection are prone to inconsistency and error; as ever, it is up to the researcher to be alert to these shortcomings and to exercise appropriate scepticism and caution in their use and reporting.

What types of question were used?

Question design is also crucial. In many large-scale surveys a great deal of resources are invested in designing and developing the questions. While this might eliminate or reduce the potential for leading or biased questions, it is still useful to cast a sceptical eye over the types of question that are being asked and to consider their relevance and fit with your own study. This is particularly important in research conducted by or on behalf of advocacy groups, the media and so on. For example, in surveys commissioned by the media, questions may mirror a headline or newspaper rhetoric or may contain limited options for responses and so potentially bias any results (Finney and Peach 2004). In the days following the 9/11 terrorist attacks, public opinion polls conducted in the United States quickly identified Osama Bin Laden as the country's leading enemy. However, by the first anniversary of the attacks, polls were revealing that the majority of Americans believed that it was Saddam Hussein who was personally responsible. The popular view for this shift of opinions is that it was a consequence of the Bush administration's publicity campaign to prepare the American public for the war in Iraq. This, according to Althaus and Largio (2004: 1) is a 'misperception' – the key reason for the shift in opinion was the wording and format of the poll questions that were being asked. Their review of opinion polls taken in the weeks and years following the attacks suggests that in open-ended

questions asked during September 2001, few Americans held Saddam Hussein responsible, but this changed when respondents were given a choice of possible perpetrators in forced-choice questions. The universal switch to forced-choice questions after 2001 served only to compound the misperception (Althaus and Largio 2004).

How relevant are the data to your own research questions?

Once you have selected your data and, to the extent that it is possible to do so, determined its purpose and objectivity, the next thing to think about is how closely the data matches your own research questions and empirical aims. Crucial to this is the fit between the variables that you are interested in studying as part of your own research design and the variables that actually exist in the dataset. Ensuring congruence between variables can be an important challenge for the secondary analyst.

Do the variables match?

One example of the challenges faced by researchers when there is a potential mismatch between one's research questions and the variables available for secondary analysis became apparent in our 2005 secondary analysis of the *PISA 2000* study (White and Smith 2005). The context for this study was current national and international concerns about shortages in teacher supply and retention (for example, OECD 2002). *The PISA 2000 School Questionnaire* gathered the views of school principals about, among other things, their experience of teacher shortages and teacher turnover and the impact this had on student learning. Whereas our interest was in examining the school level factors which impact on teacher shortages and turnover; the questions asked in *PISA 2000* referred to 'teacher shortage/inadequacy' and 'teacher turnover' and school principals' *perceptions* of the extent to which either of these two phenomena 'hindered' the learning of 15-year-old students in their schools. There were several problems with the data generated by *PISA 2000* and their fit with our research questions. First, 'teacher shortage/inadequacy' conflates two different (and perhaps unrelated) phenomena. It is impossible to discern from individual responses whether a school principal's answer relates to shortage, inadequacy or a combination of the two. Second, it is unclear how a school principal would be able to assess whether any of these problems 'hindered' the learning of students in their institutions. This task would be difficult to address in a dedicated research project and may, arguably, be impossible outside a controlled intervention. In view of this, school principals' responses to these questions cannot easily be taken at face value. Finally, of course, these data only relate to school principals' *perceptions* of shortages and turnover, no actual data on teacher vacancies was collected or asked for in the *PISA*

2000 study. Therefore, for the purpose of our study, several conceptual adjustments had to be made. The most fundamental was to accept that the school principals' responses represented perceptions of whether a problem with teacher shortage/inadequacy or turnover exists in their institution. If school principals were worried about teacher supply or quality, it may be reasonable to expect that they would demonstrate this concern in a questionnaire addressing these issues. Therefore in our analysis, the responses were treated as proxies for a general concern regarding teacher shortages/inadequacies.

Even so, the data provided by PISA were arguably the best quality available. The data provided a unique opportunity to examine the views of a large number of school principals from a range of education contexts, making this study a new development on the small-scale work which often characterises research in this area. Our concerns about the match between research questions and variables were explicitly documented in reports emerging from this project and our warrant paid due regard to its limitations (White and Smith 2005).

Do your definitions match?

Another important consideration is the extent to which your definition of a variable matches the definition from the agency which collected the data. Although definitions are vitally important, they can be very vague and very complex. Consider the definitions of 'further study' and 'assumed to be unemployed' which accompany the UK Higher Education Statistics Agency (HESA) data on student destinations at the end of higher education:

> [Further study only] includes those who gave their employment circumstances as temporarily sick or unable to work/looking after the home or family, not employed but not looking for employment, further study or training, or something else and who were also either in full-time or part-time study, training or research, plus those who were due to start a job within the next month or unemployed and looking for employment, further study or training and who were also in full-time study, training or research. (HESA 2007)

While 'assumed to be unemployed':

> includes those students who gave their employment circumstances as unemployed and looking for employment, further study or training, and who were also either in part-time study, training or research or not studying, plus those who were due to start a job within the next month and who were also in part-time study, training or research or not studying. (HESA 2007)

Perhaps it is the (mis)use of punctuation, a desire for precision, or the need to include some reference to further study, but these are not particularly clear and straightforward definitions. Contrast HESA's 'assumed to be unemployed' with the International Labour Organisation's (ILO) definition of unemployment; this is the most widely used definition of unemployment and is the one adopted by the UK government. Although in its fullest form it is quite long and complicated, it is readily summarised as:

> All persons above a specified age who during the reference period were
>
> a. 'without work', i.e. were not in paid employment or self-employment,
> b. 'currently available for work', i.e. were available for employment or self-employment during the reference period; and
> c. 'seeking work', i.e. had taken specific steps in a specified time period to seek paid employment or self-employment. The specific steps may include registration at a public or private employment exchange; application to employers; checking at worksites, farms, factory gates, market or other assembly places; placing or answering newspaper advertisements; seeking assistance of friends or relatives; looking for land, building, machinery or equipment to establish own enterprise; arranging for financial resources; applying for permits and licences, etc. (Rodda 2005)

In short, the 'unemployed' are those people who 'have not worked more than one hour during the short reference period, generally the previous week or day, but who are available for and actively seeking work' (O'Higgins 1997: 1).

In addition to overly complex definitions, the researcher also needs to ensure that the definition that they are working with matches that used in the secondary dataset. For example, does your definition of unemployment match the standard ILO definition? Definitions for the variables used in surveys and other secondary sources are usually to be found in the documents that accompany the data or in the appendices of reports and other publications.

Do you have the resources to retrieve and analyse the data?

Despite the obvious advantages of secondary analysis in terms of economies of time, money and personnel, the scale and complexity of many large studies means that it can take time to develop the skills and expertise needed to use the datasets effectively. For example, working out exactly what the variables represent and how they are coded, cleaning the dataset, deleting variables that may be of less interest, combining variables and recoding variables, all takes a lot of time. The

time one needs to commit to preparing and carrying out a secondary analysis is not always realised by other researchers. For example, a colleague and I recently requested a very small amount of pump-priming money to fund an exploratory analysis of the Pupil-Level Annual School Census (PLASC). Although the funding was awarded, the proposal's reviewers did remark that the amount of time requested to carry out the study (6 days) seemed rather extended. The zipped files containing the datasets ran to 22 different documents, one with over 7 million cases. The amount of work required to decide which datafiles were relevant to the study, to clean and retrieving variables, to merge datasets and so on, was considerable and that was before we could begin the analysis.

Therefore, once you have located your secondary sources, it is worth pausing to consider the extent to which the data need re-analysis. Many of the large survey, census and administrative datasets that are described in Appendix 1 will already have been extensively analysed. If you are interested in descriptive data, then they are likely already to have been presented elsewhere and there is no need to download and re-analyse the data further. For example, the publications that accompany the PISA dataset already contain extensive analysis of trends between groups of students with regard to their responses to different questions in different educational contexts. In the UK, the findings from the large national surveys are presented in the range of publications published by the Office for National Statistics: for example, *Social Trends* and *Living in Britain*. This type of data is ideal for researchers who wish to use secondary sources to present a context to more in-depth work or who have limited experience or interest in downloading, preparing and re-analysing some of the other datasets described in Appendix 1. So if you are interested in describing trends in the performance of boys and girls in international literacy tests or reviewing patterns of household expenditure, this kind of data is usually already available in aggregate form (i.e. already analysed) and just needs to be reported and correctly cited. (See Chapter 5 for some worked examples which use aggregate data of this type.)

By the same token, re-analysing data that have already been analysed is not necessarily a waste of time. For novice researchers, it is an invaluable way of becoming familiar with the range of data available as well as with the techniques of data preparation, analysis and presentation: skills that have applications in social science research more widely. Re-analysis can also lend new and original perspectives to existing data, for example, through the use of new statistical techniques or different theoretical frameworks.

Retrieving datasets can be very straightforward. Data from surveys such as PISA and TIMSS can be downloaded directly from the project websites. But access to many other datasets requires the user to register

through their institution or for their institution to be part of a wider consortium of data sharers. For example, access to the aggregate data from the UK National Censuses can be obtained via CASWEB provided the user has an ATHENS password. Access to other datasets may require registration and some international datasets can only be accessed if one's institution is a member of a data-sharing consortium, such as the Inter-University Consortium for Political and Social Research (ICPSR). Many archives have certain requirements as to the accessibility, use and storage of their datasets and it is important to read through the terms and conditions of data release and confidentiality.

The internet has made it possible to access data from sources that would previously have been impossible, or at least impracticable, for anyone other than the best connected researcher. Sitting at your desk in the English Midlands it is possible to re-analyse the data from classic American sociological studies, to examine progress towards millennium development goals in sub-Saharan Africa and review immigration trends in the Canadian provinces. However, accessing international datasets can be more complicated that it might seem at first, particularly when the data are presented in a language with which the researcher is unfamiliar. Unsurprisingly, many international archives, although they have English-language versions of their websites, make their datasets available only in the national language.

In deciding on your data, you should also consider whether you have the technical skills to analyse the full dataset. Many datasets require a certain amount of technical expertise and a familiarity with concepts such as statistical weightings and so on. However, fortunately, there are a number of courses available to help researchers with even the most limited skills in data management and analysis. The teaching datasets that accompany several of the large government surveys are also a good place to practice managing and analysing data. Secondary analysis of these large datasets need not be the preserve of skilled social statisticians; with a degree of patience and perseverance even a novice researcher will find much to reward their endeavours in the huge range of secondary sources that are available for analysis. A selection of UK-based training opportunities are introduced in Appendix 1 but it is worth reminding the the reader of the huge range of courses available at the University of Essex research methods summer school, the Cathy Marsh centre at the andtheinstitutionsandcentresassociatedwith
the ESRC Research Methods hub based at the University of Southampton.

Are the data of good quality?

Official statistics categories 'occupy contested terrain, the numbers they contain are threatened by misunderstanding as well as self-interest' (Porter 1995: 41).

Assessing the quality of your selected dataset is crucial. Indeed one of the problems with secondary analysis is that errors that may have been present in the original data may no longer be visible (Kiecolt and Nathan 1985). The difficulty, of course, is in knowing what the errors might be and how they might be remedied. However, the cautions that apply when examining data for errors and checking that they are measuring exactly what you expect them to be measuring are no different those you would apply when assessing the accuracy of any other piece of research.

Another crucial consideration when analysing secondary data is the possibility that the indicators adopted either by the secondary analyst or the original researcher(s) may have tenuous connections to the concepts under study: 'Slippage between concept and indicator is an ever present danger in secondary analysis' (Hyman 1972: 23). One example of this is the PISA indicator of parental wealth. While this indicator ostensibly measures 'wealth' by asking about ownership of consumer items, it does not correlate very highly with other indicators including parental occupation (Gorard and Smith 2004). It is also worth considering whether any reviews or commentaries have already been written about the data. This might unearth existing analyses that may complement your own research but could also reveal methodological shortcomings. Barretta-Herman's re-analysis of the IASSW World Census 2000 points to several limitations of the study, including a lack of clarity in terminology used and a lack of specificity in the aims of the study (Barretta-Herman 2006). (See Chapter 2 for a discussion about shortcomings in international comparative tests.)

What are the sampling strategies and response rates?

Two important questions for when you have identified the dataset are: how was the sample drawn and is it sufficiently representative to allow generalisations to be made to the wider population? Many large-scale surveys employ rigorous sampling techniques to try to ensure that the data that they collect are representative and will support generalisations. For example, in the PISA studies, a sample of schools is drawn from all schools in the participating country. Schools are selected on criteria such as size, to ensure an even spread of different types of institution. For each school sampled, additional replacement schools will also be selected. These replacement schools share the key characteristics of the main sample schools and are substituted for the main schools in the event of

Table 4.1 School response rates, before and after replacement *PISA 2003*, selected countries

Country	Participation rate before replacement	Participation rate after replacement
Canada	80	84
Finland	97	100
France	89	89
Germany	98	99
Japan	87	96
Korea	96	100
Spain	98	100
UK	64	77
USA	65	68

Source: OECD 2005b: 171–172

their non-participation. However, even such rigorous sampling techniques cannot always ensure complete and representative coverage of the population. Table 4.1 shows the response rates for a selection of countries that participated in *PISA 2003*. The first column of figures shows the response rates for schools from the original sample and the second column the rates after replacement schools had been approached. For many countries, the response rates are very robust. In Finland, Korea and Spain almost all the schools selected to be part of the original sample participated in the survey. The use of replacement schools ensured that coverage in these three countries was 100%. On the other hand, the UK and the USA are the two countries with the lowest school response rates even after replacement schools had been contacted. Indeed, these are the lowest rates for all 41 countries participating in *PISA 2003*. The poor response rates for the UK and USA are too low for the findings to be generalised to the larger country population and any results for these countries should be used with caution or even disregarded.

Other studies may also suffer from response rates far lower than the desired 100%. For example, the Youth Cohort Study has seen a steady decrease in its response rate – from over 70% in the late 1980s to 47% in 2004 (Table 4.2). The response rate for the various sweeps of the Labour Force Survey tends to be in the region of 63% (Higgins 2007). Similarly, the response rate for the 2003 Young People's Social Attitudes survey was 66% (Park et al. 2004). The 2000 International Association of Schools of Social Work (IASSW) World Census generated a response rate of only 21% and that was heavily skewed in favour of certain regions (Barretta-Herman 2006).

Response rates for surveys can be found in the technical manuals or appendices that accompany the reports and should always be consulted

Table 4.2 Response rates for the Youth Cohort Survey

Survey year	Initial used sample	Sweep 1 response rate
1985	12,180	69
1986	19,565	74
1987	21,032	77
1989	20,000	71
1991	20,060	72
1992	36,292	69
1994	27,139	66
1996	24,500	65
1998	22,500	65
2000	25,000	55
2002	30,000	56
2004	30,000	47

Note: the survey is not annual.

Source: DfES 2005

before you proceed with your analysis or investigation of the data. One of the decisions the secondary analyst has to make is whether or not they feel that these response rates are sufficiently robust to enable further analysis; this, as with many other things in social research, is a matter of judgement.

Linked to response rate is dropout, particularly in longitudinal cohort studies. One problem with analysing cohort and other longitudinal studies is the absence of substantial amounts of relevant data, often arising through participants dropping out. A cohort study like the 1970 Birth Cohort Study (BCS) uses a group of neonates and seeks permission to follow them through their lives. This study started with 16,695 cases in Britain. By 1999, 2608 were untraced, 246 confirmed emigrated, 109 died and 338 refused, leaving 13,394 cases (Bynner et al. 2000: 31). Unfortunately, the cases dropping out at each 'sweep' are not random, so introducing a substantial bias for subsequent analysis. This potential for bias should be highlighted and taken into account by analysts and their users. For example, Croxford (2006) provides an excellent summary of some of the major problems faced when conducting an analysis over time using a cohort study (in this case the Youth Cohort Study).

How timely are the data?

Another question to ask yourself is when were the data collected? And, additionally, are the data still relevant for today? Data that have been around for a long time are not necessarily of less value to researchers than data collected very recently. One potential limitation to the use of

government surveys is their timeliness: they are often at least 2 years old before they are made available for secondary analysis. An exception is the Labour Force Survey whose database is available for analysis within 14 weeks of the data collection period (Arber 2001). However, this general availability does mean that the surveys can be easily downloaded for secondary analysis but, equally importantly, the findings can also be published ready analysed in aggregate form.

It is also worth bearing in mind how relevant data collected some time ago are when applied to contemporary research questions. For example, the 1958 National Child Development Study gathered data on the attitudes and experiences of school-aged children when they were 16 years of age. These children, who were born in the late 1950s, have lived through a very different educational era from young people currently in school, most notably the introduction of comprehensive schooling in the mid-1960s, the raising of the school leaving age to 16 in 1972, the abandonment of the 11-plus examinations in many parts of the country and, of course, the introduction of the National Curriculum in England and Wales in 1988. The lessons for contemporary education research from any secondary analysis of this data are necessarily limited in their application. This is not in itself problematic; it all depends on one's research questions: as a comparative study against subsequent cohorts for example, such data would be invaluable.

Who was the information collected from?

When reviewing your dataset it is important to investigate who the actual information was collected from. In particular it is worth deciding whether the respondents would actually be in a position to answer the question with any degree of accuracy or whether their response was coloured by their own experiences, prejudices or expectations. Our research into school principals' perceptions about teacher shortages, which was described earlier, is a good example of this (White and Smith 2005). No data were collected about *actual* numbers of vacancies and turnover rates and we have only the school principals' *perceptions* about the extent to which a problem actually existed. Again, this is not necessarily problematic but it needs to be recognised in the warrant one attaches to the research. This is also important in studies which collect and then categorise data from different groups. For example, in studies concerned with participation in post-compulsory education, it may not always be clear whether the classification of occupation should be that of the potential student or of their parents. It would seem unreasonable perhaps to base the occupational classification of a student on their own work history when they may never have been anything other than a full-time student. But where the occupations of the two parents differ, which

is to be preferred? If one or more of the parents has not lived with the student, does this make a difference? What about more mature students? Should their occupational classification be based on the previous occupation of their parents? Should we use two different classification systems for younger and older students? If so, when should the cutoff point be? (Gorard et al. 2007). There are no straightforward answers to these questions, but the issues they raise need to be considered when you select variables for analysis and when findings are reported.

What categories are used to group the data?

Another important consideration is the type of categories that have been used to group the data. An obvious example of this comes with the categorisation of ethnicity or occupational class, which themselves reveal longstanding issues of classification (Lambert 2002; Lee 2003). The categories themselves are somewhat arbitrary and they interact importantly with each other and with other categories such as sex (Gorard et al. 2007). A further problem comes when examining trends in social categories over time as the variables collected, or the coding used, may also change over time. Consequently, it is often difficult to make genuine and straightforward comparisons over time or between groups. This is true, for example, of the Higher Education Statistic Agency (HESA) datasets in recording the ethnic origin of students in HE. Until 2001/02 there was only one category for 'white' students in the UK. Now a distinction has been made between white, white-British, white-Irish, white-Scottish, Irish traveller and other white. There are now, also, categories for a number of mixed ethnic groups, including mixed white. While this may reflect changes in society, and could increase the completion rate for this question, it makes comparison over time more difficult. The categorisation of socioeconomic groups and young people with special educational needs are two other examples.

How precise are the data?

The secondary analyst needs to be aware of potential issues in the way in which data, in particular aggregate, or summary, secondary data, are presented. For example, the use of rankings can overemphasise differences that are in fact rather small and the inclusion of too many decimal places can suggest an accuracy that is not warranted by the measure being presented. In order to be alert to such specious accuracies, the secondary data analyst will need to pay careful scrutiny to the footnotes and appendices that accompany the data and keep an eye out for guides as to the precision of the data, such as error bars and confidence intervals.

For example, results from international tests are usually presented as

Table 4.3 Progress in International Reading Literacy Study, average scores and standard errors

Country	Average scale score	Standard error
Bulgaria	550	3.8
Canada*	544	2.4
Czech Republic	537	2.3
England*	553	3.4
Germany	539	1.9
Hungary	543	2.2
Italy	541	2.4
Latvia	545	2.3
Lithuania*	543	2.6
Netherlands*	554	2.5
Sweden	561	2.2
United States*	542	3.8

*These countries all have queries next to their response rates.

Source: Mullis et al. 2003: 26

mean scores, often accompanied by a 95% confidence limit (which provide an estimate of the variability of the scores). The size of these bands means that the scores for some countries overlap and that simple ranking of countries can be unhelpful and disguise closely ranked performances. Table 4.3 shows the average scores for a selection of the highest ranking countries in the 2001 Progress in International Reading Literacy Study (PIRLS). Notice that when the standard errors of the mean score are accounted for the rankings in the table are fairly meaningless. This does not prevent much being made of the results, particularly among the media. For example, England's third place success in *PIRLS 2001* was attributed to a return to traditional teaching methods in primary schools (*Daily Mail*, 7 April 2003), the National Literacy Strategy (*DfES Press Release*, 8 April 2003, *Guardian*, 9 April 2003) and Harry Potter (*Daily Express*, 9 April 2003). Whereas the standard error indicates England could be as high as second or as low as fifth.

In checking the precision of the data it is also important to check whether the data distinguish between the groups of interest. The level of differentiation between different ethnic categories is an obvious example of this.

Who is missing from the data?

In other words: are there any groups who have been excluded from the data? This is an important consideration for secondary analysts who will need to consider carefully the nature of any information that might have

been omitted during the data collection process. It is not unusual to find that individuals are simply missing from official statistics, a situation made more problematic by not knowing who or how many people are missing. The recent debate on immigration statistics is a good example of this (BBC 2007b; *Guardian* 2007). In England at present, the Pupil-Level Annual School Census does not collect any data on the eligibility of permanently excluded students to receive free school meals, neither does it record their National Curriculum year group. This means that if you wished to study the profile of an institution in terms of individual students who receive free school meals, you would have to omit from the study all students who had been permanently excluded – as students in receipt of free school meals are one of the least successful groups in school, in terms of aggregate examination performance, excluding this group from your analysis leaves a potential bias in the data.

There is one study currently underway in a UK university which is examining the participation of men in higher and further education: this is a group that has apparently not benefited from the current widening participation agenda. But the study is only focusing on men currently undertaking access or foundation courses: that is, men who are *already participating* in education. What about the men who are not participating and who are arguably not benefiting from the widening participation agenda – surely these are the respondents who the researchers ought to be focusing on, surely these are the ones who are best placed to tell us about the barriers to participation in education? This sort of bias in research design is one that the secondary analyst, and arguably all researchers and reviewers, need to be aware of. The research is funded by the ESRC and its datasets will presumably be archived for use by future secondary analysts. This, unfortunately, is not an unusual omission in educational research (Gorard and Smith 2007) and it is certainly not a concern that should only occupy secondary analysts.

Are there any missing data?

An even more common problem for large-scale datasets lies in data missing even from existing cases. These 'missing' data, which can include 'not known', 'information refused', 'information not yet sought', and 'other' non-completed, often account for a large proportion of the responses. Missing data are a particular concern when the data request information on an individual's ethnicity and occupational group. For example, in the 2005 ethnicity data for first-year UK domiciled undergraduate and postgraduate students reported by the Higher Education Statistics Agency (HESA), the ethnic group was missing for around 10% of cases (HESA 2006). Similarly, analysis of the data for applicants to teacher training courses in England and Wales in 2005 showed that other

than 'white', 'missing' is officially the largest ethnic group (GTTR 2006). In fact, the unknown cases considerably outnumbered all minority ethnic groups combined.

Often the number of respondents identified as belonging to a minority ethnic group is quite small, leading to the volatility of small numbers when analysing trends over time or differences between groups. Similar issues concern data reported for occupational group. The data presented by UCAS on the occupational group of students applying for and being accepted to undergraduate programmes in British universities consistently reveal around 20% of cases whose social group is unknown (UCAS 2006). This has important implications now that the UCAS will pass these data directly to HEIs as part of widening participation initiatives (BBC 2007c). Of course, we have no way of knowing the occupational group of those students whose data are missing. Consequently, the high proportion of missing cases in an analysis using these variables could significantly bias the results being presented, even where the overall response rate is high. This means that any differences over time and place or between social or ethnic groups, needs to be robust enough to overcome this bias (Gorard et al. 2007).

In our recent secondary analysis of training statistics for initial teacher training in England, we found that around 8% of trainees failed to complete their postgraduate teacher training programme (Smith and Gorard 2007). The reasons recorded for why an individual might not complete their course do vary and are considered to be extremely important for those who monitor participation in teacher training and particularly the use of financial incentives. But the data on reasons for leaving training are missing for over half of the trainees with 'unknown' reasons given for a further 6% (Table 4.4). This leaves reasonable data for only around one-third of trainees and makes reliable evidence for the reasons trainees fail to complete initial teacher training impossible to discern from these data.

Although secondary data analysis has an important strength in enabling researchers to research small and hard-to-reach groups, care does need to be taken to ensure that sample sizes are robust and representative. For example, Connor (2001) used the Youth Cohort Study to try and identify a sample of students who had achieved the required grades but did not continue to higher education. In total, 600 such potential participants were identified, but the achieved sample size for the study was only 176 (29%). The problems in identifying those who opt out completely from post-compulsory education was further highlighted in this study when it emerged that 36% of the students previously thought to be non-participants had actually returned to education, possibly after taking a year out. This meant that this study was able to identify only 63 out of a possible 600 students who so far as the

Table 4.4 Reasons for leaving ITT courses

	N	%
Academic failure	374	4
Transferred to other HEI	21	0.2
Health reasons	186	2
Death	12	0.1
Financial reasons	59	0.7
Other personal reasons	2307	27
Written off after lapse of time	49	0.6
Exclusion	18	0.2
Gone into employment	98	1
Other	426	5
Unknown	482	6
Missing	4450	52
Total	**8482**	**100**

Source: Smith and Gorard 2007

researchers could tell did not participate in higher education in spite of the fact that they had achieved the required grades.

Ethical considerations when using secondary data

One advantage of secondary data analysis is that it doesn't require the researcher to collect new data. In practical terms, it means that it is not necessary to go through the many steps that are increasingly required in order to obtain ethical approval for research, a particular advantage for undergraduate and masters'-level dissertation study. But this does not mean that research involving secondary data analysis is necessarily free of ethical consideration. In particular, there is the notion of informed consent and the problems with using data for a purpose other than that for which they were collected and for which the respondent did not necessarily agree. Although survey data and other secondary data may be anonymised, it does not mean that the moral obligations that would hold for any researcher gathering ethnographic or interview data are absent when secondary data are used. Surveys in particular involve an interviewer entering the home and perhaps asking sensitive questions and establishing a rapport of trust with the respondent. However, the advantage of structured interviews, in this respect, is that the respondent can choose not to answer questions (Dale et al. 1988).

A somewhat broader ethical consideration concerns the type of information that is collected in research, in particular in government-sponsored surveys. Here the questions asked in these surveys reflect issues of political as well as contemporary interest. As Neuman (2003)

argues, official statistics are 'social and political products' (p. 328), and such assumptions guide the data collection and categorisation process, and dictate which data we collect. In this way, the collection of official statistics brings new attention to an issue that might not have existed before. But of course, these same processes also dictate *any* data that researchers will collect. While using secondary data can bypass some of the ethical considerations that preoccupy researchers conducing more in-depth work, this does not mean that ethical issues are not relevant. Rather, the secondary researcher has an ethical responsibility to respect the data in their possession and not to misuse them. Of course, on the ethical plus side, secondary data analysis is an unobtrusive research method and its use ensures that no further intrusion into the homes and lives of the respondent is required.

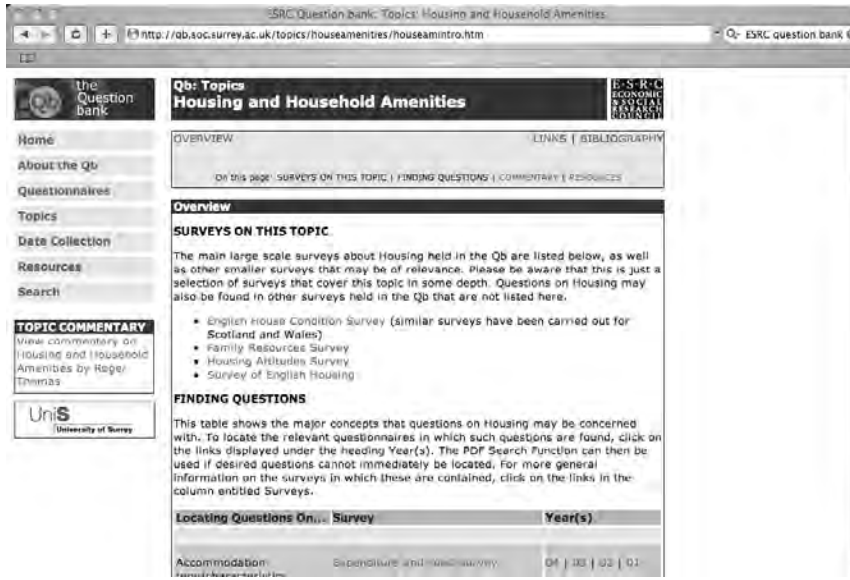
Using the 2005 General Household Survey to examine patterns in the ownership of consumer durables

This section applies the questions we have discussed in the first part of this chapter to a worked example which uses the 2005 General Household Survey to examine the ownership of consumer durable items. We begin by describing how to locate a suitable dataset before considering practical issues such as data collection, sampling and response rates and the management of variables. The questions, definitions and variables that are shown here all come from the files which accompany the GHS dataset.

Locating a dataset

A good place to start your search for a dataset is with the ESRC Question Bank (see Appendix 1 for more details). It is possible to search the Question Bank by looking for a particular survey, or as in the example here, by searching through the topic menu. Using the topic menu leads you to an alphabetic list of topics, clicking on the 'Housing and Household Amenities' link takes you to a list of four potential surveys that contain questions pertaining to the ownership of household durables (Slide 4.1, scroll down to see household durables section). These are the British Household Panel Study, the General Household Survey, the Expenditure and Food Survey and the Family Resources Survey. The example here looks at data from the General Household Survey.

Once you have selected the study on which you will base your secondary analysis, the Question Bank provides links to further information which, in the case of the General Household Survey, is available from the Office for National Statistics and the UK Data Archive. Having located a



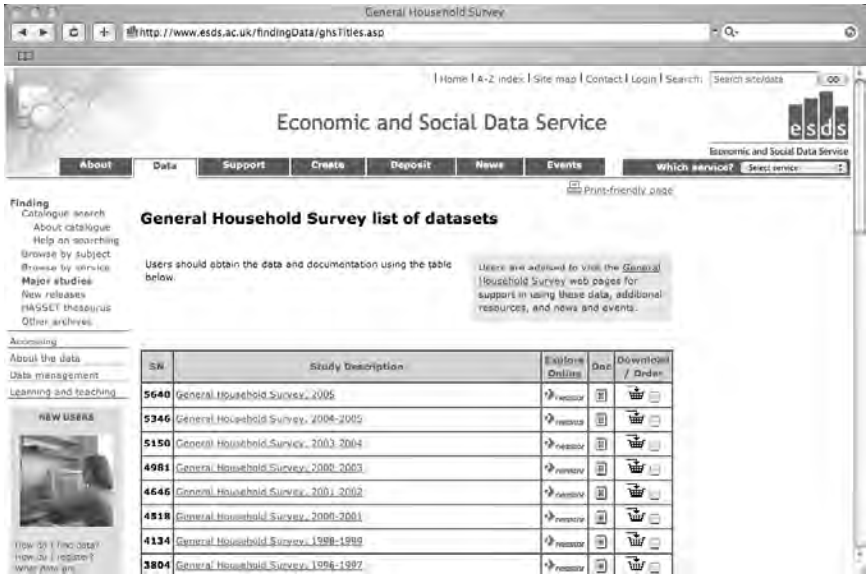
Slide 4.1 ESRC Question bank: Housing and Household Amenities databank

potential dataset, the next step is to find out about the background to the survey: its aims, its sponsors and procedures for data collection, sampling and so on. If you are interesting in looking at the aggregate data for this survey (that is data that have already been analysed and that are presented in publications such as *Living in Britain*) then background details of this nature will appear in the publication's appendices.

For those working with the raw data, more detailed information is available when you download the documentation associated with a particular survey. In the case of the General Household Survey (GHS), you will need to log onto the Economic and Social Data Service website (see Appendices 1 and 3 for an introduction to this facility and guidance on setting up an account) and download the data files and documentation associated with the GHS (Slide 4.2).

There are around a dozen files associated with the 2005 version of the GHS. Downloading these files will provide you with all the information you need in order to assess the suitability of using the GHS to help answer your research questions. The documents available for download include the GHS questionnaire, summary report, response rates, as well as the actual dataset (see later for more details).

If you decide to download the data yourself, a useful place to start is with the GHS teaching datasets. Further details about the range of support available for users of the GHS are introduced in Appendix 1. Otherwise have a look at the ESDS GHS website (ESDS 2007a, 2007b).



Slide 4.2 Downloading the General Household Survey datasets from ESDS

When you download the 2005 GHS datasets you will receive a large file containing folders which have the following documents:

- Two datafiles in the format of your choice (e.g. SPSS or Stata), one containing the household dataset only, the other the combined household and individual (client) data.
- GHS 2005 Overview Report.
- GHS 2005 Appendix A: Definitions and Terms.
- GHS 2005 Appendix B: Sample Design and Response.
- GHS 2005 Appendix C: Sampling Errors.
- GHS 2005 Appendix D: Weighting and Grossing.
- GHS 2005 Appendix E: Questionnaires and Show Cards.
- GHS 2005 Appendix F: Summary of Main Topics Included in GHS Questionnaires 1971–2005.
- EXCEL file containing Table of Questionnaire Changes 2004–2005.
- GHS 2005 Coding Frames.
- GHS 2005 Derived Variable Specifications.

The following section will use these documents to introduce the GHS and take you through the steps needed to prepare it for analysis. Although the discussion is focused mainly on the steps needed in order to analyse raw data from the GHS datasets, an understanding of variable names, questionnaire items, response rates and so on, is still important

even if you are looking at the summary data from publications such as *Living in Britain*.

Background to the survey

The General Household Survey (GHS) is a continuous national survey of people living in private households, conducted annually by the Office for National Statistics (ONS). The main aim of the survey is to collect data on a range of core topics, covering household, family and individual information (ESDS 2007a, 2007b). It is sponsored by several government departments including the Department of Health, the Department for Work and Pensions and the Scottish Executive. The main GHS comprises a household questionnaire completed by the household reference person (see later for a definition) and an individual questionnaire completed by other members of the household who are aged 16 and over.

Data collection

Information for the 2005 GHS was collected by face-to-face interview with trained interviewers using computer-assisted personal interviewing (CAPI). Interviews were sought with all members of the sampled household aged 16 and over; proxy information for children was also obtained. To help maximise response rates for the GHS, a letter was sent in advance of an interviewer calling at an address. The letter briefly described the purpose and nature of the survey and prepares the recipient for a visit by the interviewer (Office for National Statistics 2007).

Sampling and response rates

In GHS 2005 16,560 addresses were sampled. The GHS aims to interview all adults aged 16 or over at every household at the sampled address. It uses a probability, stratified two-stage sample design. The main sample is drawn from postcode sectors, which are similar in size to wards and the secondary sampling units are addresses within those sectors (Office for National Statistics 2007: 1). Table 4.5 shows the outcome of visits to the addresses selected for the 2005 survey. Out of the 18,695 addresses that were selected, 17,184 were eligible and this produced a sample of 17,310 eligible households (as some addresses contained more than one household). Interviews (including proxy interviews) were carried out with every member of 11,980 households. In a further 291 households, interviews were conducted with some, but not all, members of the household. This produced a total of 12,271 full or partial interviews (Office for National Statistics 2007: 6).

Table 4.5 Sample of addresses and households, GHS 2005

Selected addresses	18,695
Ineligible addresses	
Demolished or derelict	1511
Used wholly for business purposes	
Empty	
Institutions	
Other ineligible	
No sample selected at address	
Address not traced	
Eligible addresses	17,184
Number of households at eligible addresses	17,310
Number of households where all individual interviews achieved (including proxies)	11,980
Number of households where some but not all individual interviews achieved	291

2005 data include last quarter of 2004/5 data due to survey change from financial year to calendar year.

Source: Office for National Statistics 2007: 7, Table B2

Variables and questionnaire items

The variables of interest in this example are ownership of consumer durables such as a home computer, a washing machine, a car, a colour television and a telephone. As well as selected household characteristics, namely, employment status, occupational group, gross annual income and number of school-aged children in the household. The best place to find out how these variables are measured is in the questionnaire that is available as an appendix in the documentation accompanying the data-set. In the GHS the questionnaire is available as *Appendix E: Questionnaires and Show Cards*. This document describes the questions and prompts that the interviewer used to elicit information from the respondents. Box 4.1 shows the questions associated with ownership of a home computer; similar questions were used for the other variables linked with consumer durables, although the car ownership questions are more detailed. The items in bold are the variable names as they appear in the datafile as this enables one to know exactly which question the variable is linked to.

Box 4.1 Use of home computer, GHS 2005

Now I'd like to ask you about various household items you may have – this gives us an indication of how living standards are changing.

Does your household have any of the following items in your (part of the) accommodation?

INCLUDE ITEMS STORED OR UNDER REPAIR.

INCLUDE ITEMS OWNED, RENTED OR ON LOAN.

IF ANY MEMBER POSSESSES AN ITEM, THE HOUSEHOLD POSSESSES IT.

Ask all households

47 Computer Home computer?

EXCLUDE: VIDEO GAMES

Yes 1
No 2

Ask if household does not have a home computer

(Computer = 2)

48 CompWhy (You said your household doesn't have a computer). Is that because you . . .

don't want one 1
would like one but cannot afford it 2
or is there some other reason? 3

Source: Office for National Statistics 2007: 10–12

The questions here are retrieved from the section of the household questionnaire which focuses on consumer durables. Extracts from the script are shown in Box 4.1, beginning with the interviewer reading a preamble.

A similar structure was used for questions relating to other consumer durables.

Definitions of variables

As indicated earlier, there are two datasets available for the 2005 GHS. The first contains information relating to the household and is derived from interviews with the household reference person (file name **ghs05_client_hhld**) and the second contains data derived from interviews with other household members (here labelled the **ghs05_client** file). A household is defined as: 'a single person or a group of people who have the address as their only or main residence and who either share one meal a day or share the living accommodation' (McCrossan 1991, cited in Office for National Statistics 2007: 6).

Similarly, the household reference person (HRP) is defined as the following:

- in households with a sole householder that person is the household reference person
- in households with joint householders the person with the highest income is taken as the household reference person
- if both householders have exactly the same income, the older is taken as the household reference person (Office for National Statistics 2007: 7).

In the Household database, the data that relate to an individual (for example, their economic activity or ethnicity) is linked only to the HRP. This might be problematic for variables such as ethnicity as the HRP may belong to a different ethnic group to other members of the household. If you are interested in the ethnic group of separate members of the household, then it is better to use the Client dataset.

In this example the following variables were used to elicit background information on the respondents.

Occupational group

Some of the background information that is presented for analysis in the GHS datafile was not collected directly from the respondents. Instead, the variable is a composite of several different questions which were then put together to make a new (or derived) variable. One example is questions on occupational group where several questions about the nature of the respondent's job are combined to give one variable. This single variable is often all that is presented in the analysis files. From April 2001 the National Statistics Socio-economic Classification (NS-SEC) was introduced for all official statistics and surveys, which replaced previous classifications of social class and means that comparisons with data categorised using older occupational and social categories has to be discontinued. The GHS Household file presents three different versions of the NS-SEC classifications. For example, there is a three-class version (**hrpsec3**) comprising:

- managerial and professional occupations
- intermediate occupations
- routine and manual occupations.

There is also a five-group (**hrpsec5**) or an eight-group classification (**hrpsec8**) comprising:

- large employers and higher managerial occupations
- higher professional occupations
- lower managerial and professional occupations
- intermediate occupations
- small employers and own account workers
- lower supervisory and technical occupations
- semi-routine occupations
- routine occupations
- never worked and long-term unemployed.

The number of categories that you use depends on the level of complexity you wish to add to your data and also the substantive and theoretical interest that you bring to your research questions.

School-aged children

In the Household questionnaire, the HRP is asked to list all the people living in the household. This information is used to produce a variable for the number of school-aged children in the household, as a question about school-aged children does not actually appear in the household questionnaire. This variable is derived from other information on the questionnaire, such as number and age of dependent children and appears as variable (**schagech**) in the main datafile.

Employment status

This is presented in the datafile as the variable **hrpilo** and labelled as 'economic status'. It is a derived variable that has data in the following categories:

- not available, economic status not known
- child
- working (including unpaid)
- government scheme with employment
- government scheme at college
- unemployed (ILO definition)
- other unemployed
- permanently unable to work
- retired
- keeping house

- student
- other inactive.

Definitions for these variables are available in a separate appendix (Office for National Statistics 2007). For example, the GHS uses the International Labour Organisation (ILO) definition of unemployment. This classifies anyone as unemployed if:

> [H]e or she was out of work and had looked for work in the four weeks before interview, or would have but for temporary sickness or injury, and was available to start work in the two weeks after interview. (Office for National Statistics 2007: 4)

Household income

In the worked example shown here, household income is represented by the *usual gross weekly household income*. In addition, a range of variables are also given in the main dataset and represent both the actual and grouped weekly income for the HRP (and partner). The definition of total income for an individual refers to:

> [I]ncome at the time of the interview, and is obtained by summing the components of earnings, benefits, pensions, dividends, interest and other regular payments. Gross weekly income of employees and those on benefits is calculated if interest and dividends are the only components missing. If the last pay packet/cheque was unusual, for example in including holiday pay in advance or a tax refund, the respondent is asked for usual pay. No account is taken of whether a job is temporary or permanent. Payments made less than weekly are divided by the number of weeks covered to obtain a weekly figure. Usual gross weekly household income is the sum of usual gross weekly income for all adults in the household. (Office for National Statistics 2007: 9)

Missing data

Missing data are also an important concern and it is up to the secondary analyst to judge whether the amount of missing data jeopardises the reliability of the results. If you are working from the main dataset, the easiest way to find out how many data are missing is to run a simple frequency calculation. (If you were to analyse this in SPSS you would need to select the following sequence of commands starting on the main toolbar with **Analyse** → **Descriptive Statistics** → **Frequency** and then placing the variable of interest into the main box.) For example, for the variable gross weekly income, data were missing for a range of reasons from around 11% of households (Table 4.6).

Table 4.6 Missing responses for gross weekly income variable

	N	%
Not available	975	8
Data refused income	370	3
Data received	11,457	89
Total	**12,802**	**100**

Source: Office for National Statistics 2007

Summary

This chapter has introduced some of the practical considerations for selecting a dataset for secondary analysis. These were then illustrated by a brief worked example, which assessed the suitability of the General Household Survey for a secondary analysis of ownership of consumer durables. Of course, the key determinant of whether a dataset is suitable for secondary analysis is its fit with your research questions. Here I have tried to demonstrate some practical steps for accessing the dataset, examining the questionnaire for question wording and definitions of variables, as well as examining response rates, sampling strategies and missing information. The majority of the information needed to undertake this assessment is available in the reports that accompany the raw datasets or, in the case of aggregate data, this can usually be found in the appendices that are attached to the publication. The final decision regarding suitability of the dataset, however, lies with the researcher and their assessment of the fit between dataset and research questions.

Part II

Part II will apply what we have learnt about locating and using secondary data by taking the reader through a series of worked examples. The aim of this is twofold. The first is to show the reader the wealth and diversity of secondary sources that can be used to answer a series of substantive research questions. The second aim is to introduce some of the data management techniques that can be helpful when analysing very large and complex datasets. Chapter 5 provides a number of short examples that use aggregate secondary data, that is, data that have already been analysed and are now presented in summary form. The examples used in Chapter 5 include an introduction to using the data management tools: CASWEB and NOMIS; the use of summary data from several government surveys; as well as government administrative data to reflect social inequalities. Chapter 6 describes the secondary analysis of raw data from two large-scale British surveys: the Youth Cohort Study and the British Social Attitudes Survey. While Chapter 7 has a more international focus, with its first example using administrative secondary data produced by California and its second providing an exploratory analysis of the PISA study.

Chapters 6 and 7 place a primary emphasis on techniques that are useful for managing large-scale datasets, in other words, they seek to encourage effective data husbandry by guiding the user through the stages needed to prepare data for analysis. Techniques that are described include accessing datasets and deleting un-needed cases and variables. The methods of analysing the data used in Chapter 6 are largely descriptive, while in Chapter 7 two regression models are developed. However, the focus on methods of analysis in these two chapters is secondary – there are many excellent books to guide novice and experienced reader alike through the arithmetic and statistical highlights of analysing numeric data. The following chapters simply present an

introduction to managing the data prior to analysis, with supplementary guides to data management given in Appendices 2 and 3. Despite these relatively modest aims, effective data management or data husbandry is not to be overlooked – it is the foundation of good secondary analysis. Spending time preparing and familiarising yourself with the data rather than rushing headlong into analysis can help avoid the abstracted empiricism, button pushing or data-dredging habits that are so often criticised in numeric research.

There are several ways in which the reader may wish to use the chapters in Part II. They may be read as one might read the chapters in a conventional book: some of the substantive findings to emerge from the worked examples are actually quite interesting. Alternatively, the reader might wish to download the data and follow through the stages in preparation and analysis themselves. If you do decide to take this approach, then I suggest that you use the website that accompanies this book to help locate the datasets. Finally, the chapters may be used as a reference text when using specific datasets or management techniques. For example, if you are interested in using the British Social Attitudes Survey, a worked example is given in Chapter 6 or if you wish to download PISA, instructions appear in Chapter 7.