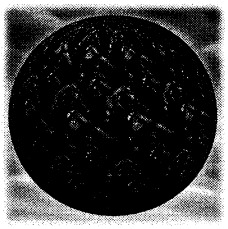

Data Collection, Primary vs. Secondary

Joop J. Hox
Utrecht University, Utrecht, The Netherlands

Hennie R. Boeije
Utrecht University, Utrecht, The Netherlands

Glossary

codebook A description of the variables, the possible values, and their location in the data file. In many cases, there is a second section detailing the original studies methods and procedures.
data archive An organization that acquires, archives, and disseminates data for secondary research.
data structure The organization of the data. This can be simple (e.g., rectangular data matrices in which rows are objects and columns are cases) or complex (e.g., hierarchical and relational databases).
meta-information The description of a data file, usually in the form of a codebook.
primary data Original data collected for a specific research goal.
qualitative data Data involving understandings of the complexity, detail, and context of the research subject, often consisting of texts, such as interview transcripts and field notes, or audiovisual material.
quantitative data Data that can be described numerically in terms of objects, variables, and their values.
secondary data Data originally collected for a different purpose and reused for another research question.

Data collection, primary vs. secondary, explains the advantages and disadvantages of collecting primary data for a specific study and reusing research material that was originally collected for a different purpose than the study at hand. After a brief discussion of the major data collection strategies in primary research, we discuss search strategies for finding useful secondary data, problems associated with retrieving these data, and methodological criteria that are applied to evaluate the quality of the secondary data.

Introduction

To collect data, social scientists make use of a number of different data collection strategies. First, experiments and quasi-experiments are important because they typically involve a research design that allows strong causal inferences. Second, surveys using structured questionnaires are another important data collection strategy because they typically involve collecting data on a large number of variables from a large and representative sample of respondents. Third, within a qualitative research design the data collection strategy typically involves collecting a large amount of data on a rather small, purposive sample, using techniques such as in-depth interviews, participant observation, or focus groups.

Primary data are data that are collected for the specific research problem at hand, using procedures that fit the research problem best. On every occasion that primary data are collected, new data are added to the existing store of social knowledge. Increasingly, this material created by other researchers is made available for reuse by the general research community; it is then called secondary data. Data may be used for (1) the description of contemporary and historical attributes, (2) comparative research or replication of the original research, (3) reanalysis (asking new questions of the data that were not originally addressed), (4) research design and methodological advancement, and (5) teaching and learning.

Data sets collected by university-based researchers are often archived by data archives; these are organizations set up chiefly for the purpose of releasing and disseminating secondary data to the general research community. In addition to universities, other organizations such as

national and regional statistical institutions also collect data. More and more, such data are also archived and made available, usually via the statistical agency itself. Finally, universities, research institutes, and individual researchers themselves may decide to make their data available. For data archives and, to a smaller extent, the large statistical agencies, the systematic preservation and dissemination of data are part of their institutional mission. For individual researchers and small research units, this is not the case. Nevertheless, the Internet holds several interesting Web sites where individuals or research units offer access to their data sets.

Most of the secondary data sets contain quantitative data; that is, the information consists of studied objects whose characteristics are coded in variables that have a range of possible values. A qualitative database consists of documents, audiocassette or videocassette tapes, or the transcripts of these tapes. Increasingly, qualitative researchers share their data for secondary analysis.

Social scientists who intend to study a particular theoretical problem or a specific policy issue have the choice to collect their own data or to search for existing data relevant to the problem at hand. The most important advantage of collecting one's own data is that the operationalization of the theoretical constructs, the research design, and data collection strategy can be tailored to the research question, which ensures that the study is coherent and that the information collected indeed helps to resolve the problem. The most important disadvantage of collecting one's own data is that it is costly and time-consuming. If relevant information on the research topic is accessible, reusing it gains benefits. The data can serve to answer the newly formulated research question, smooth the pilot stage of a project, or provide the researcher with a wider sample base for testing interpretations at far less cost and with greater speed.

This nicely sums up the advantages and disadvantages of using secondary data. The disadvantage is that the data were originally collected for a different purpose and therefore may not be optimal for the research problem under consideration or, in the case of qualitative data, may not be easy to interpret without explicit information on the informants and the context; the advantage of using secondary data is a far lower cost and faster access to relevant information.

Primary Data Collection

The Experiment

One major primary data collection strategy is the experiment. In an experiment, the researcher has full control over who participates in the experiment. The researcher manipulates one or more independent variables following a planned design and observes the effects of the independent variables on the dependent variable, the outcome variable. The essence of an experiment is that the research situation is one created by the researcher. This permits strong control over the design and the procedure, and as a result the outcome of an experiment permits causal interpretation. This is referred to as the internal validity—the degree to which the experimental design excludes alternative explanations of the experiment's results. At the same time, the fact that the experimenter creates the research situation, often in a laboratory setting, implies that the situation is to some degree artificial. The problem here is the ecological validity—the extent to which we can generalize the results of our study to real-life situations. Experimental laboratory studies put emphasis on those variables that are easily manageable rather than on variables that reflect the everyday activities of people coping with real-life situations. Typically, because an experiment involves setting up experimental situations and exposing subjects to different stimuli, experiments involve a relatively small number of subjects and variables. However, because there is strong control over the design, most experiments make an effort to manipulate several variables, using designs that permit conclusions about both their individual and their combined effects. Several handbooks describe a variety of experimental designs, including designs for longitudinal research and case studies.

The Social Survey

A second established primary data collection strategy is the interview survey. In a survey, a large and representative sample of an explicitly defined target population is interviewed. Characteristically, a large number of standardized questions are asked and the responses are coded in standardized answer categories. A survey is carried out when researchers are interested in collecting data on the observations, attitudes, feelings, experiences, or opinions of a population. Information on subjective phenomena can be collected only by asking respondents about these. Surveys are also used to collect information about behavior. Behavior could in principle be studied by observation, but this is often expensive, or impossible, as in questions about past behavior. Social surveys are usually targeted at a household population, but they can also aim at a specific subpopulation. A specific form of survey is interviewing key informants from communities or organizations. These individuals are purposively selected because they are formal or informal nodes of information and therefore in a position to provide the researcher with informative responses to the survey questions or to point the researcher to other sources of information. The strong point of survey research is that it can provide information about subjective and

objective characteristics of a population. The major methodological problems in interview surveys are obtaining a representative sample and the validity of the responses given by the respondents. Obtaining a representative sample is usually accomplished by drawing a random sample from the population, using scientific sampling methods. However, in most Western countries survey nonresponse is considerable and increasing, which may threaten the representativeness of the sample. In addition, both respondent and question characteristics can affect the responses. To ensure valid responses, interview questions must be carefully designed, evaluated, and tested.

Qualitative Research

Qualitative researchers examine how people learn about and make sense of themselves and others and how they structure and give meaning to their daily lives. Therefore, methods of data collection are used that are flexible and sensitive to the social context. A popular method of data collection is the qualitative interview in which interviewees are given the floor to talk about their experiences, views, and so on. Instead of a rigidly standardized instrument, interview guides are used with a range of topics or themes that can be adjusted during the study. Another widely used method is participant observation, which generally refers to methods of generating data that involve researchers immersing themselves in a research setting and systematically observing interactions, events, and so on. Other well-known methods of qualitative data collection are the use of focus (guided-discussion) groups, documents, photographs, film, and video.

Settings, events, or interviewees are purposively sampled, which means guided by the researcher's need for information. Provisional analyses constantly change this need, and therefore sampling takes place during the research and is interchanged with data collection. Contrary to probability sampling, which is based on the notion that the sample will mathematically represent subgroups of the larger population, purposive sampling is aimed at constructing a sample that is meaningful theoretically; it builds in certain characteristics or conditions that help to develop and test findings and explanations. Sampling strategies include aiming at maximum variation, snowball sampling, critical case, and stratified purposeful.

The intense role of the researcher brings about issues with regard to reliability and validity. That the researchers are their own instrument is necessary to gain valid knowledge about experiences or the culture of a specific individual or group; to reduce the reactivity of the research subjects, prolonged engagement is recommended. Another issue is the lack of control over the researchers' activities; therefore, researchers should keep detailed notes of their fieldwork and the choices they make in order to increase replication and reproducibility. Many other quality procedures have been developed, such as triangulation, member checks, peer debriefing, and external audits.

Solicited and Spontaneous Data

A distinction that involves all primary data collection techniques is that between data that are solicited and data that are spontaneous. In experiments, surveys, and much qualitative research, the researcher uses a stimulus (experimental variable, survey question, or open question) to elicit information from the research subjects. Explicitly soliciting information has the advantage that the researcher can design the data collection to optimally provide data given the research question. However, the disadvantage is that the research subjects are aware that they are taking part in a scientific study. As a consequence, they may react to the announcement of the study topic, the institution that sponsors the study or carries it out, the individual experimenter or interviewer, and so on. It is not clear whether the recorded behavior or response is the "true" behavior, that is, whether it is the same behavior that would have occurred naturally, if it had not been elicited.

The possible reactivity of research subjects can be circumvented by observing natural activities or the traces they leave behind, without disturbing the research subjects in any way. Nonreactive or nonintrusive primary data collection methods include (covert) observation and monitoring. Observation, which can be done in the actual location or remotely using video technology, can lead to both quantitative and qualitative data. Increasingly, technological advances make it possible to monitor activities without disturbing the subjects. For instance, media research in general no longer relies on a panel of respondents who report on their television viewing; instead, in selected households a monitoring device is installed in the television that monitors the television use and transmits the information to the researchers without disturbing the respondents. Scanning devices are used to monitor consumer behavior. Internet behavior can also be monitored. For instance, when people visit a specific Web site, it is simple to monitor which banners and buttons they click on, how long they stay on the page, where they come from, and where they go to when they leave the site. All this provides information without directly involving the subjects in any way.

Summary of Primary Data

Table 1 lists the types of data in primary data collection. The list is not intended to be exhaustive; rather, it is indicative of the primary data collected in social research.

Table 1 Examples of Primary Data in Social Research

	Solicited	*Spontaneous*
Quantitative	Experiment Interview survey Mail survey Structured diary Web survey	(Passive) observation Monitoring Administrative records (e.g., statistical records, databases, Internet archives)
Qualitative	Open interview Focus group Unstructured diary	(Participant) observation Existing records (e.g., ego-documents, images, sounds, news archives)

Secondary Data Collection

For some social research questions, it is possible to use data collected earlier by other researchers or for other purposes than research, such as official statistics, administrative records, or other accounts kept routinely by organizations. By virtue of being archived and made available, any type of primary data can serve as secondary data.

Using secondary data presents researchers with a number of characteristic problems. First, researchers must locate data sources that may be useful given their own research problem. Second, they must be able to retrieve the relevant data. Third, it is important to evaluate how well the data meet the quality requirements of the current research and the methodological criteria of good scientific practice.

Finding Useful Secondary Data: Search Strategy

The main sources of information are the official data archives. These organizations are established for the precise purpose of acquiring, archiving, and disseminating data for secondary research. For example, the described aim of the Qualitative Data Service (Qualidata) at the U.K. Data Archive is providing research, learning and teaching communities with significant real-life data that can be reanalyzed, reworked, and compared with contemporary data and that will, in time, form part of the cultural heritage as historical resources. Nowadays, archives usually store electronic data. Important data archives maintain Web sites that describe the data in the archive and coordinate information about the existence of available sources of data wherever they are housed. These catalogs usually contain information on the subject of the original study, the data collection mode used, the number of variables, and the number of subjects.

After a data set that looks attractive has been located, the next step is to obtain a more detailed description of the study. Well-documented data sets come with a detailed description of the methods and procedures used to collect the data. For example, if the original study is a large-scale survey conducted by an official statistical agency, the detailed description of the study will describe the interview procedures, the sampling design, whether sampling weights are to be used, and any transformations performed by the original researchers on the data set. In addition, the text of the questionnaire will be available. All this information may be needed to judge whether the secondary data suit the present research problem.

In addition to official data archives, many other data sets are available. The problem here is twofold: tracing them and judging their usefulness. National and regional statistical agencies sometimes make data available for secondary analysis. Also, several universities and research institutes offer data files, and there are individual researchers who maintain Web sites about topics of special interests. However, data that are not part of an official data archive are typically poorly documented. This makes it much more difficult to evaluate whether they are useful and whether their methodological quality is good enough so that the researcher can trust the results of the original analysis.

The Internet is a vast but extremely disorganized source of information. To use it effectively, we must be able to formulate a search strategy using appropriate keywords. Some Internet search machines support logical expressions, which allows searching for combinations of keywords. Because no individual search engine covers all of the Internet, the best approach is to use a metasearch program such as Copernic. Metasearch programs send a search request to a collection of search engines, combine the results into one list, and sort this list on the basis of estimated relevance. If relevant secondary data files are available on the Internet, such a procedure has a reasonable chance of finding them.

The projected ability to assess the data quality is part of the search strategy. Before the methodological quality is assessed (which may mean retrieving the codebook and sometimes even the data), potential secondary data sources must be screened for practicality or feasibility. The criterion of practicality means that the data must be promising because they appear to cover the topic (contents and variables), are described in a language that the researcher can read, and are readily obtainable. The availability of meta-information, that is, information that describes the data and thus enables methodological quality assessment, is an important requirement in selecting potentially interesting secondary data sets.

Retrieving Secondary Data

Secondary data contained in an official data archive are easy to retrieve. They may be available as files obtained via

the Internet or as data files on CD-ROM or DVD data discs. Because of the widespread use of email and the World Wide Web, it has become easier to retrieve data in a usable format (i.e., in a form that can be input simply into the appropriate analysis software). Different archives set different conditions for delivering data to individual researchers, and these terms may also differ according to the specific data set asked for. Some data sets are freely available, without cost. Others may be expensive, either because the original data collector wants to recover part of the expense of the primary data collection or because of the cost of offering the existing data sets in a reusable format for new research purposes. Conditions can sometimes involve the use of the data itself. In most cases, use of the data is restricted to the researchers who request them. Therefore, if secondary data are needed for teaching purposes, and therefore must be supplied to classes of students, special permission for this use is usually needed. In some cases, the only requirement is that the original study and the data archive be cited; in others, the original investigator must be informed of the purpose of the secondary analysis and must give explicit permission for the data to be reused for that purpose. However, for the majority of secondary data sets in official archives it is straightforward to obtain them at a reasonable cost.

If data sets are found that are not part of an official archive, locating the original investigator might be part of the research process. Even if the data are simple to obtain, locating information about the original data collection procedures may be difficult. Nevertheless, if the secondary data are to be used for research, this additional information is vital to evaluating the methodological quality of the data. Again, it may be necessary to locate the original investigator or research agency in the hope that a research report is available that describes the methods and procedures of the original study.

If data are stored in an archive, they are usually available in several different formats. The lowest common denominator is a text or ASCII file, meaning the numeric values are available in a raw data file without additional information. Usually, the data are organized as a rectangular data matrix, with fixed columns for the variables and one or more lines (records) per subject. In addition, there usually is a codebook, which describes the variables in the data set, what values they may legitimately take, and their location in the raw data file. The recipients of the data can use this information to input the data into the statistical package they use. In many cases, the data are also available in the format of the major statistical packages such as SPSS or SAS, so the variables already have appropriate names, the variable values are labeled, and missing data indicators are set.

The rectangular data structure is the simplest structure. Other data structures are hierarchical and relational data structures. In a hierarchical data structure, multiple hierarchical sources of information are combined. For example, an educational study may have data both on pupils and on their schools, with a number of pupils belonging to the same school. Such data can exist as two separate files, linked by a school identification variable, or as a single file, containing both pupil and school variables. In a relational data structure, there are several files that are connected on the basis of a predefined variable or structure. For instance, an educational study might have data on pupils and schools, but also data on the pupils, their families, and the neighborhoods they live in. This does not fit a tidy hierarchical structure, but all the data are related through the pupils.

Although it is common practice to analyze survey material collected for other researcher's projects, earlier qualitative research data are considered to be less of a source and are scarcely used, except by social historians. Sharing qualitative data generates mixed feelings that have to do with practical, methodological, and ethical issues. The value of secondary qualitative data is questioned because much of the meaning is supposed to be lost when, for instance, tone of voice and interviewees' gestures are ignored. There is also concern that the original fieldwork context cannot be recovered, which is necessary for adequate interpretation and understanding. Ethical issues have to do with confidentiality and obtaining consent for the preservation and future reuse of the raw materials by those other than the original researcher.

Qualitative data can be of all kinds. Open interviews generally result in long transcriptions, which can be stored as a text file. Again, the lowest common denominator is the simple text or ASCII file. However, qualitative data can contain many different information sources, such as video tapes and audiocassette tapes, photographs, diaries, and so on. The most valuable qualitative data sets for future reanalysis are likely to have three qualities: (1) the interviewees have been chosen on a convincing sampling bases, (2) the interviews follow a life-story form rather than focusing on the researcher's immediate themes, and (3) recontact is not ruled out.

Evaluating the Methodological Quality of Secondary Data

If the secondary data set stems from previous research, it should be considered that each research project is undertaken with a view to answering a specific set of research questions. Choices are made regarding which information is going to be collected, how sampling will proceed, and specifically how the information is collected. In quantitative research, these generally involve deciding on a general strategy of data collection—carrying out a survey and subsequently deciding on specific survey instruments or

scales to be used and a specific mode of surveying respondents, such as a face-to-face survey. In qualitative research, similar decisions are made about the general strategy—carrying out a qualitative interview study and subsequently on the sample, instruments, data analysis, and structure of the report. In both cases, (small) differences in definitions, classifications, and concepts among studies may have an impact on the reusability of the data.

How well the choices and limitations of the original study fit the current research problem is a matter of judgment. Secondary researchers must consider carefully whether the data appropriately fit their research question. In general, it is acceptable if the limitations of the available data limit the secondary analysis to some extent, for instance, by impeding tests of some specific hypothesis. However, it is not acceptable if the limitations of the data make it necessary to change the research question in order to be able to use them at all.

In any case, when secondary data are used it is very important to evaluate the quality of the data closely. This requires that additional information be archived with the data itself. Such information should at least include the study purpose, operationalization, data collection details (who, when, and where), entities being studied and sampling criteria, and any known biases. Because searching for suitable secondary data and retrieving them in usable form can be time-consuming, an integral part of the search strategy is to make certain that such additional metadata describing the data and their quality are in fact available. This is also important because the data sets deemed sufficiently interesting to be archived are usually very large. Typically, secondary researchers need to select specific subsets of variables or groups of cases. Some providers of secondary data offer a proprietary interface to select subsets from the data; others provide a huge data file from which the customers must select what they need. Again, all this demands a good description of the data.

For the purpose of protecting the confidentiality of the respondents, the data that are made available are sometimes purposively altered. When data are available on a microlevel, combining information from different sources may make it possible to identify individual respondents. For instance, by combining information on job (librarian), age (42), gender (female), and town (San Diego), it may be possible to identify a specific respondent with reasonable probability. For reasons of disclosure control, secondary data providers sometimes transform the data by adding random numbers from a specified statistical distribution. This makes it impossible to identify individual respondents. And, obviously, it will also impede the analysis; however, statistical procedures exist to obtain correct results when the data have been deliberately contaminated this way.

The analysis of quantitative data often starts with a process in which the data are cleaned—incomplete records may be edited or automatically imputed, nonnormal data may be transformed, aggregated scores may be calculated, and so on. This is all based on the assumptions and interpretations of the primary researcher. Researchers who reanalyze the data for another research purpose may not always agree with the assumptions implied in the data cleaning and coding. Instead, they may prefer to use other methods that are more in line with their own research purpose. If the secondary data contain all the original variables and if the codebook documents all data manipulations, researchers can apply their own data-cleaning process. However, tracing the preliminary data-cleaning process can be a difficult and very time-consuming process. At the very least, researchers should be alerted and aware of changes to and recodings of the original raw data when secondary data are used.

If meta-information on the secondary data is incomplete or even totally lacking, it becomes impossible to assess the reliability and validity of the original procedures. Such data, however, can still be useful for teaching purposes, for use as example data or as practice data for students. In fact, there are several data archives and other organizations that make data available specifically for teaching purposes; some of these are general, whereas others aim at a specific analysis technique.

See Also the Following Articles

Focus Groups • Interviews • Surveys

Further Reading

American Statistical Association. http://www.amstat.org/

Berg, B. L. (2001). *Qualitative Research Methods for the Social Sciences.* Allyn & Bacon, Boston.

British Library National Sound Archive. Oral history collection. http://www.cadensa.bl.uk

Copernic. Available at: http://www.copernic.com/

Council of European Social Science Data Archives (CESSDA). http://www.nsd.uib.no/cessda/

Cresswell, J. W. (1998). *Qualitative Inquiry and Research Design: Choosing among Five Traditions.* Sage, Thousand Oaks, CA.

Dbmscopy. Available at: http://www.dataflux.com

De Leeuw, E. D., and de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In *Survey Nonresponse* (R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, eds.), pp. 41–54. Wiley, New York.

De Vaus, D. (2001). *Research Design in Social Research.* Sage, Thousand Oaks, CA.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (2002). *Confidentiality, Disclosure and Data Access: Theory and*

Practical Applications for Statistical Agencies. Elsevier, Amsterdam.
Experimental-data Archive. http://www.radcliffe.edu/murray/data
Fink, A. (1998). *Conducting Research Literature Reviews.* Sage, Thousand Oaks, CA.
Hammersley, M., and Atkinson, P. (1995). *Ethnograph: Principles in Practice.* Routledge, London.
International Consortium for Political and Social Research (ICPSR). http://www.icpsr.umich.edu/
Kerlinger, F. N., and Lee, H. L. (2000). *Foundations of Behavioral Research.* Harcourt, Fort Worth, TX.
Library of Statistical Problems. http://lib.stat.cmu.edu/DASL/
Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.) (1997). *Survey Measurement and Process Quality.* John Wiley, New York.
Maxwell, J. A. (1996). *Qualitative Research Design: An Interactive Approach.* Sage, Thousand Oaks, CA.
Maxwell, S. E., and Delaney, H. D. (1990). *Designing Experiments and Analyzing Data.* Brooks/Cole, Pacific Grove, CA.
Qualidata. Qualicat. http://www.qualidata.essex.ac.uk/search/qualicat.asp
Seale, C. (1999). *The Quality of Qualitative Research.* Sage, London.
Silverman, D. (2000). *Doing Qualitative Research: A Practical Handbook.* Sage, London.
StatTransfer. Available at: http://www.stattransfer.com
Stewart, D. W., and Kamins, M. A. (1993). *Secondary Research: Information Sources and Methods.* Sage, Newbury Park, CA.
Thompson, P. (2000). Re-using qualitative research data: A personal account. Forum: *Qual. Soc. Res.* [Online Journal], **1**(3), Available at: http://qualitative-research.net/fqs/fqs-eng.htm
UK Data Archive (UKDA). http://www.data-archive.ac.uk/
Weiss, R. S. (1994). *Learning from Strangers: The Art and Method of Qualitative Interview Studies.* Free Press, New York.