some of the more recent developments regarding sample comparability, pooling samples, calculating significance tests using multiple samples, and so on.

Researchers of all social science disciplines in both academic and applied settings will find this text extremely useful. The chapter on locating data provides practical information for all secondary analysts, from advanced undergraduates to seasoned researchers. The chapter on making effective use of data introduces analysis problems that all secondary users confront at one time or another. While making complete use of the solutions proposed may require additional reading and more statistical training than some readers will have, the material is presented in a manner that will provide all readers with a clear understanding of how to approach previously collected survey material. This dual emphasis makes *Secondary Analysis of Survey Data* the most valuable guide to the secondary analysis of survey data to have come along in recent years.

—*Richard G. Niemi*
Series Co-Editor

SECONDARY ANALYSIS OF SURVEY DATA

K. JILL KIECOLT
Louisiana State University

LAURA E. NATHAN
Mills College

1. INTRODUCTION

The method of data collection known as survey research has for some time been central to investigation in the social sciences. The survey is a rather flexible tool, and survey research easily lends itself to the exploration of a wide range of topics requiring different types of data (demographic, attitudinal, behavioral, and so on). The role of survey research will undoubtedly be enlarged by the availability of large-scale social surveys which can be used for a variety of research projects in addition to the ones for which they were originally intended.

Traditionally, social scientists have been encouraged to collect their own data, regardless of the data collection method used. After selecting the question to be addressed, researchers are charged with designing their research in keeping with the problem at hand. When survey research is the method of choice, questions can be developed to elicit precisely those data that are needed.

Unfortunately, independent data collection by the individual investigator has become increasingly difficult. Constraints of the current economic climate and declining resources for research in the social sciences have made it necessary for more researchers to rely on existing survey data. The potential for accomplishing original research with precollected data is nonetheless tremendous. Further, inasmuch as original data cannot be gathered for times past, analysts of change must rely on existing data to probe shifts in attitudes and behavior. Secondary analysis is thus gaining a central role in contemporary social research.

This type of analysis is neither a specific regimen of analytic procedures nor a statistical technique. Rather, it is a set of research endeavors that use existing materials. It differs from primary research in that primary analysis involves both data collection and analysis, while secondary analysis requires the application of creative analytical techniques to data that have been amassed by others.

"Meta-analysis," or "quantitative literature review," should be differentiated from the secondary analysis of documents or surveys. Meta-analysis integrates the findings from a universe (or sample) of investigations of some phenomenon. That is, the study itself becomes the unit of analysis. Meta-analysis involves using published research results to compute an overall level of significance for an array of comparable test statistics, such as correlations or *t* values (see Rosenthal, 1984, for a review of methods). It can also be used to find the average effect size of a treatment across studies—for example, the mean difference in outcomes between control groups and groups receiving therapy (Glass, 1978). Meta-analysis has been used primarily to evaluate experimental research in psychology and education, but the technique may also be applied to research in other disciplines.

Documents are written materials that contain information. They are rarely developed with social research in mind, yet they are potentially rich sources of information on social phenomena. Document studies are often qualitative, but they may also employ quantitative content analysis techniques. Since the researcher must construct the categories for analysis, determine the recording unit, and decide on the system of enumeration, even content analysis assumes a degree of subjectivity.[1]

Documents are used extensively in historical research and often provide data over time where no similar data exist. While documents themselves may be classified as primary (i.e., eyewitness descriptions of behavior or events) or secondary (i.e., second-hand accounts), document studies are virtually always secondary analyses. Documents are used less frequently, however, than surveys for secondary analysis.

Surveys have been employed to elicit information on a wide variety of topics from both general and specialized populations. There is now an abundance of surveys on population characteristics, attitudes, and behavior. Secondary analysis of existing surveys allows researchers access to data from large, national samples—data that would be difficult for a lone researcher to gather. Both large and more specialized archives have been established that typically make surveys available for a modest fee (see Chapter 2). Technological advances, such as the effective storage of data in machine-readable form and the availability of statistical

computing programs, as well as widespread researcher access to computers, have also contributed to the popularity of using precollected survey data for original research. As a consequence, secondary analysis of survey data is likely to maintain a dominant position in social science research for the foreseeable future.

Advantages of Secondary Survey Analysis

The primary advantage of secondary survey analysis is its potential for resource savings. Secondary research requires less money, less time, and fewer personnel and is therefore attractive in times of economic fluctuations, when the funds available for research are limited or uncertain. With data already collected, the costs are only those of obtaining the data, preparing them for analysis (such as ensuring that all data are computer-ready and compatible with the system), and conducting the analysis (buying computer time, and so on). Compared with the time normally required to collect data in social research, the time necessary for acquiring an appropriate data set is miniscule. Further, a researcher can complete a research project independently, thereby eliminating the need for ancillary research staff. Secondary analysis also obviates the need for researchers to affiliate with a large organization in order to command the backing necessary for acquiring adequate survey data.

Another advantage is that secondary analysis circumvents data collection problems. Data archives furnish a large quantity of machine-readable survey data spanning many topics, time periods, and countries (see Chapter 2). Many available data sets provide the benefits of nationally representative samples, standard items, and standard indices. Both data availability and improvements in technology facilitate research. Growing numbers of researchers have access to computer facilities and computer software packages such as SPSS^{x} (1983) and SAS (1982) to simplify analysis.

A variety of research projects can be accomplished with precollected data. When used in exploratory research prior to fielding a new survey, secondary analysis can uncover aspects of a research problem that require elaboration, groups that need to be oversampled, grounds for hypothesis revision, and the need to refine and improve existing measures (Hyman, 1972). Secondary analysis may be employed for a variety of research designs, including trend, cohort, time-series, and comparative studies (see Chapter 3). Existing data can also be combined with other types of data to investigate a problem more thoroughly. For

example, they can be combined with primary data to render an analysis dynamic, or they can be used to supplement in-depth interviews. Demographic and historical studies and research conducted under time constraints, such as policy-related projects, often require the use of existing data.

Our increased familiarity with and use of preexisting data encourage social scientific progress. Data sets such as the General Social Survey and the American National Election Study are widely used, and investigators who employ these data sets can turn to other researchers with questions on data handling. The widespread use of particular data sets also allows authors greater ease of reporting. On the basis of earlier articles which have used or discussed a particular data set, most readers will be familiar with relevant aspects of the survey such as the sampling procedure and question wording.

Familiarity with existing databases also offers researchers the opportunity to build on what is available and conduct trend studies. For example, to study trends in the enjoyment of work, Glenn and Weaver (1982) commissioned the inclusion of some work enjoyment questions on the 1980 Gallup national survey which duplicated items that had not been asked since 1955. Finally, the better acquainted researchers become with existing databases, the greater the potential for creative new research. Ideas for studies often emerge from interaction between a researcher's substantive interests and his or her intimate knowledge of information contained in data files.[2]

Limitations of Secondary Analysis

Although the advantages of secondary analysis clearly outweigh the disadvantages, there are limitations. Many of the problems that secondary analysts encounter are intrinsic to the survey method, but some are unique to secondary analysis. A major problem is data availability. Despite the development of data archives, researchers sometimes have trouble locating what they need. Some topics lend themselves more readily to secondary analysis than others; for example, researchers interested in drug abuse, crime, and physical health are likely to have an easy time finding data. In more specialized areas, such as mental health epidemiology, however, there are relatively few publicly available databases, and investigators must depend on the generosity of individuals who own private data files. Often primary researchers are reluctant to share their data, as reputations are made by publishing work from a controlled body of data. A different sort of data availability problem stems from a major mismatch of primary and

secondary research objectives. Sometimes when information on specific items or individuals is desired, data are available only in scaled or aggregated form.

Some secondary analysts complain about the time involved in acquiring data sets from archives. While admittedly it often takes several weeks to obtain data, this time is short compared to the time it takes to design a survey and collect data. The situation is analogous to that of adoptive parents a decade or two ago. Many bemoaned the four- to six-month wait involved in adopting a child, despite the fact that it would take nine months to produce their own.

In order for a data set to be usable, researchers must know exactly what they are analyzing. Without complete and accurate documentation of all data on a file, it is difficult or impossible to locate needed information. Inadequate documentation sometimes occurs with data tapes produced by small research firms on a contract basis, but it is not usually a problem with data sets produced by large academic organizations. Major archives usually provide complete data documentation in a codebook that sequentially lists the variables in a data file, as well as describing them and their assigned values. Whenever possible, data files should be checked by comparing the marginal distributions of a subset of the variables with those reported by the original investigators.

Errors made in original surveys often are no longer visible, and it is impossible to differentiate interviewing, coding, and keypunching errors. Moreover, the survey procedures that were followed may not have been sufficiently documented to enable secondary analysts to appraise errors in data. Trivial sources of error, such as that from sampling design, may be magnified when a survey is put to other than its original use, and such errors may be compounded by combining surveys. For example, a study using a national sample that excludes the institutionalized population could draw misleading conclusions about very young or very old adults because of the relatively high proportions of these groups who reside in various institutions (Hyman, 1972). The problem would be compounded by using such samples for a trend study of old or young adults to the extent that the proportions of these groups in institutions have changed over time.

A related problem occurs when a secondary researcher wants to study a specific subpopulation but has only a broad-based, nationally representative sample. Unless the sample is enormous, there may be too few cases to conduct the desired statistical analysis. In such cases, pooling surveys for data on rare populations is a possibility (see Chapter 3).

Data quality is another reason that some researchers are leery of secondary research. Data files from surveys employing nationally representative samples, properly designed questionnaires, and rigorous procedures for interviewing and coding do not always exist. Even surveys of high quality may have measurement problems. Invalidity is of concern to the extent that survey items are imprecise measures of the concepts a secondary analyst has in mind, or that the variables have been poorly operationalized. Surveys rarely contain all the variables of interest to the secondary researcher, and even when they do there may be too few indicators of a concept for reliable measurement. Thus researchers sometimes need to use a number of surveys to assemble arguments that cannot be developed with the data from one survey alone. Using multiple surveys compounds potential error, however, and issues of comparability arise when measures of a concept are not strictly equivalent. In sum, secondary analysts must frequently make do with measures that are not precisely those desired. Often this results in criticism from peers for lacking hypothetically perfect indicators, or for proceeding atheoretically with research.

Another disadvantage of secondary analysis is the possible inhibition of creativity. If researchers use the same data sets repeatedly and are limited by the variables contained therein, scientific progress will be thwarted to some extent. More globally, continued use of the same indices and data sets may limit the scope of social science research. However, we believe that the inclusion of the same measures is necessary to ensure comparability. As long as new items are continually incorporated into surveys, advances will be made in the social sciences.

The increased availability of good survey data for secondary analysis is something of a mixed blessing to the degree that it has contributed to so-called "trendy" social research. That is, some researchers obtain a data set, apply a currently popular statistical technique, and then look for a problem to investigate. Without theory, however, the utility of social research is called into question. The proliferation of survey data for secondary analysis offers tremendous opportunity, but the "data set in search of analysis" approach yields only trivial findings.

2. LOCATING APPROPRIATE DATA

In defining research problems, some secondary analysts generate ideas and then search for appropriate data, while others browse through codebooks for inspiration. While both strategies are sometimes employed, a third, even more productive approach to generating ideas for research, as described in Chapter 1, involves a merging of one's general substantive interests and familiarity with existing data files. Ideas for research projects are usually influenced by one's knowledge of existing data. Prospective users often learn about the existence of pertinent data through sharing information with other researchers about data sets that bear upon their mutual interests, or through journal articles that cite data sources in the text or accompanying tables.

In the past, secondary analysts usually had to depend on identifying primary researchers who were willing to share their data, while today most secondary analysts obtain surveys from academic archives. During the late 1950s, about 20 years after the initial proliferation of surveys, data began to be systematically preserved and stored. Social scientists throughout the world recognized the broad research potential of these surveys, and with financial support from such agencies as the National Science Foundation (Geda, 1978), they worked to create archives to house and distribute machine-readable social data.

Social Science Data Archives

Social science data archives vary considerably in terms of source and level of funding. The better the funding, the greater the depth and breadth of the services provided by the archive. External funding (from agencies and foundations) has been central to the development and support of data archives, as archives typically charge only enough to cover the costs of their services. The current state of the economy and recent cutbacks in funding for education and research have slowed the development of many social science data archives. Hence some archives charge membership fees to other institutions to help maintain archive quality.

Archive size and extent of services are directly related to the level of funding. Some archives have rather limited holdings (a few major data sets or only those data sets generated by individuals who are affiliated with the larger institution), while others have literally thousands of data files. The extent of services generally increases with size.

The main service that social science data archives provide is the distribution (by loan or purchase) of machine-readable data files and accompanying documentation. Data are virtually always furnished on magnetic tape, but some data files are available on cards for users who have access only to older equipment. While some (typically smaller)