SECONDARY DATA ANALYSIS

In social science research, you may often hear the terms primary data and secondary data. Primary data is data that was collected by the researcher, or team of researchers, for the specific purpose or analysis under consideration. Here, a research team conceives of and develops a research project, collects data designed to address specific questions, and performs their own analyses of the data they collected. The people involved in the data analysis therefore are familiar with the research design and data collection process.

Secondary data analysis, however, is the use of data that was collected by someone else for some other purpose. In this case, the researcher poses questions that are addressed through the analysis of a data set that they were not involved in collecting. The data was not collected to answer the researcher's specific research questions and was instead collected for another purpose. The same data set can therefore be a primary data set to one researcher and a secondary data set to a different researcher.

Using Secondary Data

When using secondary data in an analysis, there are some important things that must be done beforehand. Since the researcher did not collect the data, he or she is usually not familiar with the data. It is important for the researcher to become familiar with the data set, including how the data was collected, what the response categories are for each question, whether or not weights need to be applied during the analysis, whether or not clusters or stratification needs to be accounted for, who the population of study was, etc. Basically, the researcher needs to become as familiar as possible with the data set and the data collection process used.

There are a great deal of secondary data resources and data sets available for sociological research, many of which are public and easily accessible. Read more about commonly used secondary data sets.

Advantages of Secondary Data Analysis

The biggest advantage of using secondary data is economics. Someone else has already collected the data, so the researcher does not have to devote money, time, energy, and other resources to this phase of research. Sometimes the secondary data set must be purchased, but the cost is almost always certainly lower than the expense of collecting a similar data set from scratch, which usually entails salaries, travel/transportation, etc. There is also a huge savings in time. Since the data is already collected and usually cleaned and stored in electronic format, the researcher can spend most of his or time analysing the data instead of getting the data ready for analysis.

A second major advantage of using secondary data is the breadth of data available. The federal government conducts numerous studies on a large, national scale that individual researchers would have a difficult time collecting. Many of these data sets are also longitudinal, meaning that the same data has been collected from the same population over several different time periods. This allows researchers to look at trends and changes of phenomena over time.

A third major advantage of using secondary data is that the data collection process is often guided by expertise and professionalism that may not be available to individual researchers or small research projects. For example, data collection for many federal data sets is often performed by staff members who specialize in certain tasks and have many years of experience in that particular area and with

that particular survey. Many smaller research projects do not have that level of expertise available, as data is usually collected by students working at a part-time or temporary job.

Disadvantages of Secondary Data Analysis

A major disadvantage of using secondary data is that it may not answer the researcher's specific research questions or contain specific information that the researcher would like to have. Or it may not have been collected in the geographic region desired, in the years desired, or the specific population that the researcher is interested in studying. Since the researcher did not collect the data, he or she has no control over what is contained in the data set. Often times this can limit the analysis or alter the original questions the researcher sought out to answer.

A related problem is that the variables may have been defined or categorized differently than the researcher would have chosen. For example, age may have been collected in categories rather than as a continuous variable, or race may be defined as "White" and "Other" instead of containing every major race category.

Another major disadvantage to using secondary data is that the researcher/analyst does not know exactly how the data collection process was done and how well it was done. The researcher is therefore not usually privy to information about how seriously the data are affected by problems such as low response rate or respondent misunderstanding of specific survey questions. Sometimes this information is readily available, as is the case with many federal data sets. However, many other secondary data sets are not accompanied by this type of information and the analyst must learn to read between the lines and consider what problems might have been encountered in the data collection process.