Social and Personality Psychology Compass 5/1 (2011): 56–75, 10.1111/j.1751-9004.2010.00329.x

A Step-by-Step Guide to Using Secondary Data for Psychological Research

Judith Pizarro Andersen[1], JoAnn Prause[2], and Roxane Cohen Silver[2]*
[1] *Cornell University*
[2] *University of California, Irvine*

Abstract

The purpose of this paper is to serve as a primer for those who have never used, or even considered using, secondary data as a resource for psychological research. Secondary data (SD) can provide a unique methodological tool with which to examine psychological issues and can serve as a valuable contribution to a program of research. However, this important resource may often be overlooked because its use can sometimes appear daunting and time-consuming. We seek to assist new users of SD by describing the process in a step-by-step manner. We address both benefits and challenges to anticipate when using SD, and discuss identifying and acquiring potential datasets, creating a personalized dataset, variable creation, statistical considerations, and the potential problem of conflicting findings when large datasets are used by multiple researchers. Our goal is to encourage researchers who are novices to the approach to consider using SD as an adjunct to their program of research.

Secondary data (SD) can provide a unique methodological resource in which to examine psychological issues. Exploring secondary data is often done in concert with other methods, such as experimental and clinical research, to provide a well-rounded examination of a psychological construct or phenomena. However, this important resource may be overlooked because its use can sometimes appear daunting and time-consuming. The purpose of this paper is to draw awareness to the use of secondary data as a valuable adjunct in a program of research, and to serve as a primer to researchers who have not previously used this methodology. Our goal is to demystify the use of secondary data by describing the process in a step-by-step manner. SD can be useful for researchers at any career stage, including graduate students searching for data for a thesis topic, junior faculty looking to augment data used in their research program, or senior researchers seeking pilot data for grant applications.

Although the terms *archival* and *secondary* data are sometimes used interchangeably in the literature, they are defined differently. *Archival* data come from examination of primary source documents such as letters, newspaper articles, or school or medical records (see, e.g., Wicke & Silver, 2009). This often requires the complex and time-consuming process of tracking down original records and transcribing these documents to create a workable dataset. The term s*econdary data* refers to data that have been collected and made available by a primary source. Secondary data are often collected for a specific purpose but can also be used to address questions in other fields of research. In addition, general repositories of data exist to aid researchers with factual statistics about a population of interest.

Benefits of Secondary Data

In thinking of ways to advance a program of research, investigators often consider learning new methodological and statistical techniques to increase their repertoire of research skills. At the same time, researchers often have limited time and resources to design and conduct large-scale data collection projects. This is one area in which SD can be quite useful. In addition, using SD also introduces multi-disciplinary perspectives into psychological research, which helps to avoid tunnel vision that is a risk within any specialization (Tomlinson-Keasey, 1996). For example, both economists and developmental psychologists use the National Longitudinal Surveys (NLS) to study the interchange between children's development, labor force economics, and a host of family circumstances.

Secondary data can arise from both cross-sectional and longitudinal research designs. While SD coming from cross-sectional designs can be helpful for understanding the prevalence of different outcomes at a slice in time (e.g., physical health or the unemployment rate), there are particular benefits of using *longitudinal* SD. Longitudinal panel data allow comparisons that are otherwise impossible in cross-sectional and trend designs. For example, longitudinal designs enable the examination of effects of employment transitions on mental health or evaluation of the long-term consequences of early age drinking (see Dooley, Prause, Ham-Rowbottom, & Emptage, 2005), or the long-term impact of a national disaster on mental and physical health (Holman et al., 2008). Longitudinal SD can also be used to explore population-specific characteristics (e.g., urban or rural environments) from currently available datasets that include decades of economic and census data (Feenberg & Miron, 1997).

Longitudinal data inform theory development by providing opportunities to posit causal explanations of events and explore the mechanisms by which processes unfold. For example, the U.S. National Study of Health and Life Experiences of Women examined alcohol use, violence and victimization, and health and lifestyle factors across a 20-year period. Using these data, Wilsnack and colleagues (Wilsnack, Kristjanson, Wilsnack, & Crosby, 2006; Wilsnack, Wilsnack, Kristijanson, & Harris, 1998) discovered that adverse childhood experiences and missing social ties were particularly significant predictors of heavier and riskier drinking patterns over time. Secondary data analyses from this impressive study can be used to explore relationships between sexual abuse and the development of alcohol use problems, clarify interactive relationships between social/cultural and individual variables, and examine multi-generational trends in alcohol dependence.

Working with SD can also provide the opportunity to examine a variety of psychological and health-related phenomena cross-culturally, as Wolfe et al. (2008) utilized to explore stigma and universal access to antiretroviral therapy for HIV in Botswana. Furthermore, multi-national and even global analyses may be made possible with SD. For example, Wilsnack, Wilsnack, Kristjanson, Vogeltanz-Holm, and Gmel (2009) have created the International Research Group on Gender and Alcohol, where researchers from forty countries are simultaneously conducting a longitudinal investigation (2007–2012) of gender, alcohol and culture. Data from these archives continue to be used to examine national and worldwide trends in alcohol use, violence, and health, among other topics.

Using SD defrays the cost and time investment necessary to design questionnaires, collect data, and maintain large, complex datasets. This benefit of SD is particularly salient for longitudinal studies where the cost to individual research groups of re-interviewing participants over time, in some cases decades, would be prohibitive. The public availability (via electronic records) of most large data repositories enables researchers at any location and level of expertise to access these resources.

Myths Regarding Secondary Data

A common misconception that may prevent the use of SD in psychological research is the concern that using it is more time consuming and complicated than other methodologies. While there exist projects of varying degrees of difficulty (e.g., analyzing lengthy longitudinal records with complex sampling weights), working with SD can and should be adapted to the skill level of the researcher. The time-consuming process of translating research constructs into usable variables is one that is undertaken whether the researcher uses SD or not; importantly, researchers using SD often have to be more creative with the available data. For example, a variable representing anger levels among inmates in a high security facility may not be available in a particular SD set, but it may be possible to create a proxy from available data on the number of angry outbursts or incident reports of violent behavior.

Another myth about using SD is that it is inferior to the alternative of collecting one's own data. Using SD is not a replacement for personal data collection; we maintain that it is most useful in conjunction with other methodologies, such as experimentation, survey research, or clinical research. Of course, it is wise to consider both the pros and cons of using SD before making a commitment to such a project.

Weighing the Pros and Cons

Disadvantages to utilizing SD include the inability to select specific questions or measurement instruments, as well as the lack of control over the precise timing of the data collection (Tomlinson-Keasey, 1996). Nonetheless, SD offer valuable advantages that should be considered.

First, SD are often collected using well-established measures with known psychometric properties. Second, many secondary datasets contain, or can be created to provide, diverse samples that are likely to be representative of more broad populations. For example, Wilsnack et al. (2008) combined two large samples (National Study of Health and Life Experiences of Women and Chicago Study of Health and Life Experiences of Women) in order to examine risk factors for and rates of hazardous drinking among heterosexual and sexual-minority women. Existing datasets did not contain sufficient samples of sexual-minority women or were based on convenience sampling with uncertain generalizability. Secondary data analyses revealed important differences between groups; specifically, sexual-minority women had significantly higher rates of hazardous drinking in adulthood and higher rates of risk factors for hazardous drinking (child sexual abuse, depression, early alcohol use) compared to heterosexual women. Finally, secondary datasets are often large enough to provide good statistical power for most types of planned analyses (e.g., see Pizarro, Silver, & Prause, 2006, where approximately 36,000 Civil War recruits were available for analyses). (For an excellent review of the pros and cons of utilizing SD, we recommend Tomlinson-Keasey, 1996; and see Jordan, 1994, for issues specific to longitudinal SD.)

Below, we provide a step-by-step guide to the use of SD in one's program of research.

Step One: Identifying SD Appropriate for One's Research Needs

Finding a secondary dataset

The first step in the process of using SD is finding the type of data needed to answer one's research questions and planning the construction of a personalized data file. It is important to identify the major study aims because the subject matter will dictate the types of data required. For example, studies of psychological functioning require the appropriate measures (e.g., depression or anxiety scales), while studies of economic stress require appropriate variables to operationalize economic stress (e.g., changes in income, unemployment, underemployment). Locating the sources of information that contain data of interest may be a multi-step process. Government funded data collections such as the U.S. Census provide aggregate data for different geographic units (e.g., census tract). Other publicly accessible datasets can be good resources for both aggregate- and individual-level information. Table 1 provides a list of potential sources of SD that may be relevant for psychologists, along with information on the original purpose for the data collection, samples included, and location of the database. Although this list is not exhaustive, it can provide researchers with some useful examples. Many of these data repositories have descriptions of their secondary datasets (e.g., codebooks and other documentation) on their web pages or provide contact information to help learn about their available resources.

Target population

It is important to consider the desired target population when selecting SD. For example, studies of social support among the elderly require a different sample composition than studies of age of alcohol drinking onset or studies that focus on children's mental health trajectories. Some samples are nationally representative while others incorporate over-sampling of certain segments of the population to insure adequate representation of under-represented groups. It is important to be aware of any limitations of the data that might affect the generalizability of one's findings. Data that are sampled in a systematic manner from the desired target population strengthen the external validity of one's conclusions.

It is also important to identify the appropriate *unit of analysis* of the target population to best address the research question. This may necessitate using SD from one or more levels including individuals, groups, geographical units (town, census tract, state), or social interactions (dyadic relationships, divorces, arrests). Secondary data may also include information from more than one source and from different types of observations (e.g., the NLSY Child/Youth survey, which is based on the children of female respondents from the NLSY79 survey). It may be necessary to merge data from different levels when linking contextual variables to individual-level outcomes (e.g., linking regional unemployment rates to the relationship between substance abuse and the probability of re-incarceration).

Temporal location of the data

Use of SD can be particularly beneficial if one is interested in the study of events that occurred during specific times in history. For example, researchers from disciplines outside psychology, government/local county agencies, and public service or non-profit agencies often collect data around specific interventions or important events. Some examples of events that require that data collection be within specific timeframes include the September 11th, 2001 terrorist attacks, the enactment of welfare laws, or the occurrence of the economic depression.

Acquiring the data and establishing contacts

Many organizations provide data they maintain, via web communication, directly to one's computer; other agencies may provide CDs or other media recordings of the data, or

Table 1 Secondary data sources and descriptions

Secondary Data Sources	Type of data	Original purpose of data collection	Sample	Notes	Location of data or URL
American National Election Studies (ANES)	Multiple studies available. Some studies (e.g., ANES 2008 Time Series Study) have repeated measures (e.g., pre and postelection measures of attitudes, etc.).	Purpose is to better understand voting, public opinion, political participation and the theoretical and empirical foundations of national election outcomes.	Multiple samples.		The American National Election Studies (ANES) http://www.electionstudies.org Click on the 'Data Center' Tab
Annual National Health and Nutrition Examination Survey (NHANES)	Interview and survey data. Exam component of medical, dental, and physiological measurements, and lab tests administered by trained medical personnel.	Started in the 1960s to examine Health and Nutritional Status among US adults and children. Questions regarding demographic, socioeconomic, dietary, and health were gathered.	Nationally representative sample (N = 5,000) per year. Children and adults.	NHANES over-samples persons 60 and older, African Americans, and Hispanics.	Centers for Disease Control (CDC, 2009) http://www.cdc.gov/nchs/nhanes/about_nhanes.htm
BBC Daily Life Survey	Daily diary survey. Respondents list all daily activities as they happen throughout the day.	The objective is to understand the UK public's time and media usage.	A population based sample every 5–10 years since the 1930s.	The 2002/3 Pioneered a technology new to market research.	The BBC Daily Life Survey. http://bbcdailylife.tns-global.com/
Block and Block Longitudinal Study (1969 – 1999)	Interviews and assessments.	A sequence of 9 independent assessments based on personality and cognitive Life, Observational, Test, and Self-report (LOTS) measures.	A 30-year longitudinal study (N = 128) of men and women age 3 years old when first surveyed.		Henry A. Murray Research Archive http://www.murray.harvard.edu/ Click on the 'Find Data' tab and search for Block
Bureau of Justice Statistics (BJS)	Multiple formats. Longitudinal and cross-sectional survey data.	US crime trends by state, age, type, perpetrator status, crime type, prosecutor's data, administrators and law enforcement statistics.	Large, representative samples.		Bureau of Justice Statistics (BJS) http://bjs.ojp.usdoj.gov/index.cfm?ty=daa#262

Table 1 (Continued)

Secondary Data Sources	Type of data	Original purpose of data collection	Sample	Notes	Location of data or URL
Centers for Disease Control (CDC)	Multiple formats. Longitudinal and cross-sectional survey data.	Population health, epidemiological data, youth, adult and family data, prenatal, lifespan, disease prevention and intervention.	Multiple samples.		Centers for Disease Control (CDC) http://www.cdc.gov/CDCForYou/researchers.html
Consumer Expenditure Survey	Repeated measures daily diary study.	Weekly (household) expenditures of frequently purchased items (food, alcohol, smoking, personal care, nonprescription drugs), as well as income and characteristics data.	Multiple samples.	Yearly data are available for purchase by CD-ROM.	United States Department of Labor http://www.bls.gov/cex/csxmicro.htm
Current Population Survey (CPS)	Survey data.	To examine employment, unemployment, earnings, hours of work, and other indicators, demographic characteristics including age, sex, race, marital status, and educational attainment, occupation, industry, and class of worker. Supplemental questions to produce estimates on topics including school enrollment, income, previous work experience, health, employee benefits, and work schedules.	Nationally representative sample of 50,000 US households. Adults 16 years of age and older.	Conducted every month for over 50 years.	Current Population Survey (CPS) http://www.census.gov/cps/
General Social Survey (GSS)	Survey data.	To examine knowledge about and attitude towards science, self-employment, Jewish identity, social inequality, terrorism preparedness, global economics, CDC high risk behaviors, sexual orientation, and clergy sex, religion, social-welfare and economic regulation, civil liberties, spending priorities, and political efficacy.	US dwelling English and Spanish speakers age 18 and over (N = 1500–2000).	The GSS uses cross-sectional and rotating panel samples.	General Social Survey (GSS) http://www.icpsr.umich.edu Click on the 'find and analyze data' tab and search for GSS

Table 1 (Continued)

Secondary Data Sources	Type of data	Original purpose of data collection	Sample	Notes	Location of data or URL
Health and Retirement Study (HRS)	Longitudinal repeated measures survey data.	Labor force participation, health transitions at end of work life, income, pension plans, health insurance, cognitive function, assets, disability, health care costs.	A sample of 22,000 midlife and older adults age 50+.	Surveys taken every 2 years since 1992.	University of Michigan Institute for Social Research. http://hrsonline.isr.umich.edu/
Henry A. Murray Research Archive	Multiple formats. Longitudinal and cross-sectional survey data.	Topics within psychology, sociology, economics and other fields. Diversity samples on race, ethnicity, sexual orientation and religion.	Multiple samples.		Harvard Institute for Quantitative Social Science http://www.murray.harvard.edu/
Inter-University Consortium for Political and Social Research (ICPSR)	Multiple formats.	Topics in: psychology, sociology, education, political science, the substance abuse and mental health archive, psychiatric epidemiology surveys, the minority data resource center, Population research in sexual-minority health, and a health and medical care archive.	Multiple samples.		The Inter-University Consortium for Political and Social Research (ICPSR) http://www.icpsr.umich.edu
National Bureau of Economic Research	Multiple formats. Longitudinal and cross-sectional survey data.	Multiple topics including, but not limited to: child health and well-being, hospital statistics, population and labor force characteristics, industry, macro and micro economics, and international trade.	Multiple samples.		National Bureau of Economic Research http://www.nber.org/data/
National Education Longitudinal Survey of 1988 (NELS:88)	Longitudinal survey data, computer-assisted personal interviews and face-to-face follow-up interviews.	School, work, and home experiences, educational resources and support, the role in education of their parents and peers, neighborhood characteristics, educational and occupational aspirations, and other student perceptions.	Nationally representative sample (N = 25,000). Eighth graders who were enrolled in public or private school in 1988.	Some achievement test data also available for eighth-graders, sophomores, or seniors.	National Center for Education Statistics (NCES), US Department of Education. NELS:88 http://nces.ed.gov/surveys/nels88

Table 1 (Continued)

Secondary Data Sources	Type of data	Original purpose of data collection	Sample	Notes	Location of data or URL
National Epidemiologic Survey on Alcohol and Related Conditions (NESARC)	Longitudinal survey data collected via computer assisted personal interviews.	The NESARC collects data on background, alcohol and drug consumption, abuse and dependence, treatment utilization, family history of alcoholism or drug abuse, tobacco use and dependence, medicine use. Current and family mental health (e.g., depression, anxiety and personality disorders, medical conditions, and victimization.	A representative sample of the non-institutionalized U.S. population 18 years of age and older.	Designed to determine the magnitude, and change over time, of alcohol use disorders and their associated disabilities.	The National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/NESARC.htm
National Longitudinal Survey (NLSY79 Child/Young Adult)	Survey and interviews. Home observation of parent-child interactions, physiological, cognitive and socio-emotional assessments.	Child-parent interaction, attitudes toward schooling, dating and friendship patterns, religious attendance, health, substance use, and home responsibilities. Parents and children ages 15 and older. Information collected: schooling, training, work experiences and expectations, health, dating, fertility and marital histories, and household composition.	Nationally representative sample (N = 12,686) of men and women aged 14–22 years old when first surveyed in 1979. In 1986, a separate survey of all children born to NLSY79 female respondents began.	Interviewed annually through 1994 and are currently interviewed on a biennial basis.	Bureau of Labor Statistics http://www.bls.gov/nls/nlsy79ch.htm
National Longitudinal Survey of Freshman (NLSF)	Longitudinal survey data.	To examine theories of minority underperformance in college Measures of family structure, neighborhood and school characteristics (age 6 &12), pre college peer networks, racial/ethnic attitudes, courses, grades, time use, living arrangements, financial matters, relationships, perceptions of prejudice on campus. Factors in the students' decision to attend college, high school SAT scores, college major, career plans, employment while in school, graduation statistics and time to degree.	Cohort (N = 3924) first-time freshmen.	Equal numbers of Asian, Black, Hispanics, and White students sampled at each of the 28 participating schools.	Office of Population Research at Princeton University Funded by the Mellon Foundation. http://nlsf.princeton.edu/

Table 1 (Continued)

Secondary Data Sources	Type of data	Original purpose of data collection	Sample	Notes	Location of data or URL
Open Access Jochen Fahrenberg Universität Freiburg Datasets	Daily diary survey and several ambulatory psychophysiological monitoring studies.	Open access data for nine different ambulatory assessment studies.	Multiple samples.	Study descriptions are in German.	Universität Freiburg http://www.jochen-fahrenberg.de/index.php?id=27
Panel Study of Income Dynamics (PSID)	Multiple formats. Longitudinal survey and extensive interview data. Telephone interview on transition to adulthood study.	Intergenerational transmission of poverty in America. Economic, health and social behavior. Original sample of 4800 American families started in 1968.	Children from these families bring sample up to 9000 US adults.	Extensive interview data on children ages 0–12.	Panel Study of Income Dynamics (PSID) http://psidonline.isr.umich.edu/Data/
Princeton Time and Affect Survey	Telephone interview. Respondents were asked to reconstruct their day (past 24 hours).	Participants reported all activities performed in a 24 period. Three activities (during non-sleep hours) were selected for more detail and ratings of emotion. Overall satisfaction with health and life ratings and questions regarding any disability that limited amount or type of work were gathered.	4000 adults reached via random digit dialing between May–August 2006.		Princeton Time and Affect Survey http://www.krueger.princeton.edu/princetonaffectand timesurvey.php
Terman Longitudinal Study of Gifted Children (Terman, 1922–1991)	Survey data.	Purpose of the data was to understand giftedness. To identify and nurture giftedness in school and across the lifespan.	Gifted youth at age 12 (N = 1528). IQs greater than 140.	The study will continue until 2020, to encompass the entire lives of the original sample of 1528 gifted youth.	Terman Longitudinal Study of Gifted Children http://www.icpsr.umich.edu Click on the 'find and analyze data' tab and search for Terman

Table 1 (Continued)

Secondary Data Sources	Type of data	Original purpose of data collection	Sample	Notes	Location of data or URL
Youth Risk Behavior Surveillance System (YRBSS)	Self-administered paper and pencil questionnaires in the classroom setting.	Responses include behaviors that contribute to unintentional injuries and violence, tobacco use, alcohol and other drug use, sexual behaviors, unintended pregnancy and sexually transmitted diseases (STDs), unhealthy dietary behaviors, and physical inactivity.	Nationally representative sample of public high and middle school students, alternative schools, juvenile justice facilities.		The Youth Risk Behavior Surveillance System (YRBSS) http://www.cdc.gov/HealthyYouth/data/surveillance.htm

This is only a sample of the secondary data resources available and is not meant to be an exhaustive list.

require registration or travel to the data site to obtain it. We strongly advise using only secondary datasets that have readily available documentation including codebooks for interpreting variable definitions and coding.

It is useful to identify other researchers who have published from the SD and are familiar with the intricacies of using that dataset. If available, examine bibliography lists to determine possible contact persons and establish contact with these individuals. They may be able to share insights about any inherent biases of the data; common fixes to problems encountered when manipulating the data; or share variables that they have cleaned and coded. This can save many hours of time when developing data files for analysis.

An example: A study of Civil War soldiers

In our own research (Pizarro et al., 2006), we sought to examine how exposure to traumatic stress impacted an individual's physical health. The limited research that had been conducted on this topic was based on sparse self-reported data, the majority of which were cross-sectional designs. Fortunately, we were able to identify a longitudinal dataset created by economists and historians where this issue could be examined: a compilation of the military and lifetime medical records of over 35,000 veterans of the American Civil War. This dataset provided a large sample of individuals who had been exposed to objective trauma events at different ages and had lifetime health records collected in a systematic manner. Although it was not designed for the specific questions we sought to answer, we believed that this SD could nonetheless be manipulated to our advantage.

We were able to obtain our identified SD directly from the organization responsible for maintaining the data (Center for Population Economics) as an electronic download. The SD that we chose was transcribed from archived records and coded by historians and economists in a manner traditional to their specific fields of research. Approaching the data from the perspective of psychologists, we relied on their careful documentation (far beyond variable names) to determine if the SD could be used to examine our constructs of interest. In addition, it was helpful to establish contact with other researchers who published using the SD to solicit their advice on our design, selection of variables, and analytic strategies. Once our research question was identified, operational definitions of our constructs were outlined. In this case, we used standardized definitions of objective trauma exposure as defined according to the Diagnostic and Statistical Manual for Mental Disorders (DSM-IV; American Psychiatric Association, 1994) (e.g., Prisoner of War experience, witnessing comrades die), and objective physical health conditions as defined by the International Classification of Diseases, 9th Revision, (ICD-9, World Health Organization, 1999). Clear operational definitions allowed us to compare our findings across studies, between research groups, and across time.

Specifically, we examined the health of individual soldiers, and compared groups of soldiers (contrasting the effect of age of exposure on lifetime health and mortality). When comparing individuals in one group (a company or regiment) to individuals in another group, we had to adjust statistically for clustering; an adjustment that accounts for the fact that soldiers within one group are more likely similar to one another (they share similar environmental exposures) than soldiers outside of their group.

For our research team, the benefits from using SD have extended far beyond the initial project. The project served as a foundation for a junior scholar's program of research, one that resulted in subsequent grant funding and publications, and an on-going collaborative partnership was formed including a junior scholar and expert senior members.

Step Two: Creating a Personalized Dataset

Organizing the project

In order to keep track of the extensive amount of information needed when utilizing SD, we recommend creating an organizational scheme from the inception of the project. We create both an electronic file system and binders with paper copies organized into the following sections: Agency or SD repository contact information; documentation including copies of codebooks, questionnaires, User Guides; articles of interest published by others who used the data; communications and notes from meetings with other researchers; personal notes, articles, and literature searches regarding the research questions; study-specific variable codebook (explained in more detail below); the number and type of analyses conducted; summaries of statistical analyses including definitions of any specific sub-samples used and any new conceptualizations of the sample across time; a record of where the data and command files for statistical software were electronically saved and backed up; and the file name of the most current version of the data and command files.

Variable extraction

Given the breadth of variables generally available in SD, it is advisable to extract a wider range of variables than might originally be thought useful to allow for unique comparisons and possible scale development or the creation of proxies (variables that similarly explain the constructs of interest). A good way to learn about the breadth of variables available in the SD is to become familiar with papers from others who have used the data, and to carefully read the data documentation including User Guides and the original questionnaires used to collect the data. For example, in our analysis of Civil War soldiers' military and medical records, in addition to gathering the number of children and wives each veteran had, we also extracted data from their birth and death certificates. We then were able to explore the association between multiple family deaths and a veteran's incidence of physical disease.

Many SD repositories provide data extraction software that makes navigation and creation of personalized datasets very straightforward. Users can often specify that the data be saved in commonly used statistical software programs. Many data extraction programs provide command files for a specific statistical software package to access the newly created data file and label, format, and assign missing values to the variables. It may also be necessary to extract data from the main SD repository in multiple sessions as the project progresses and the research evolves. As such, it is very important to include in each data extraction the variable or variables that uniquely identify observations in the SD (e.g., in our case, the veteran's military ID number) so the multiple datasets created at different times can eventually be merged.

Creating a personal variable codebook

With an organized structure in which to record the progress of the project and a data file containing the variables of interest, documentation of variables in one's personalized secondary dataset can commence. We make sure there are labels on *all* variables, as well as create very descriptive variable labels for the values of any categorical-type data that are extracted. Each data file should be completely documented by including the name of the dataset, the date it was created, a listing of all variables with labels, and missing value codes for each variable clearly defined.

A set of basic descriptive statistics should be run on each dataset to provide summary statistics such as the mean, standard deviation, minimum and maximum values, and the sample size for each variable. An important step is to carefully review these statistics to insure that variables are labeled correctly, that missing value codes are properly defined, and the sample size is correct. If summary statistics are available from the data repository, verify that the sample size and summary statistics match.

The importance of keeping a detailed variable codebook for a SD project cannot be underestimated. In our experience, working with SD can require the organization of thousands of participants; combining, creating and recoding hundreds of variables; and numerous iterations of data in order to perform the desired statistical procedures. Team members may seek to examine the data independently over the course of the project, sometimes with months between viewings. Reviewers and readers sometimes ask how variables and proxies were created. Saving communications explaining the coding and cleaning procedures of the variables enables them to be shared with other researchers, if requested. Data will likely be used for future projects or may be provided to other interested parties; either way, a clear explanation of the development of each variable is necessary. Another important reason for keeping detailed records is the accountability to science. As Freedland and Carney (1992) review, data management and accountability in behavioral and biomedical research is an absolute necessity. Keeping meticulous records helps research teams avoid unintentional data falsification and ensures that others can replicate the work.

Structuring SD data

Due to the nature of SD, it is common to work with multiple data files over the course of a project. Multiple files can be merged using statistical software programs, and if the datasets are saved in different software packages, these files can be translated to a preferred software package by using statistical transfer programs or by taking advantage of the ability of most statistical software programs to read and write files in other formats.

Once the data are in the same software program, it may be necessary to maintain differently structured datasets from which to merge information for specific analyses (rectangular, hierarchical, etc.). For example, in our Civil War soldiers SD project, two types of datasets were required. Every soldier had one military file and multiple medical pension files that contained the health records. Military files included time invariant variables such as the soldier's age and occupation at enlistment into service, number of battles in which a soldier participated, and percentage of deaths within each soldier's company. The military data file was set up in a rectangular format (wide) with each row representing a single case (veteran). Because each veteran had multiple medical visits across his life-span, the file containing the medical pension files was set up in a hierarchical format (long) where each veteran had multiple rows reflecting how many visits he made to the doctor over his post-war lifetime, and each row represented all the information collected at each visit. Information from the medical and military files was merged so that each veteran represented a single record in the merged file. Most of the commonly used statistical software programs have the ability to re-structure or re-shape the data.

Managing large datasets

With large and complex SD, multiple datasets are likely as a result of the merging of variables to create files appropriate for different research questions or statistical techniques. An efficient way to manage these datasets is to create electronic folders that represent a category of datasets (e.g., medical records, long form; military records, wide form) and place each relevant dataset in that folder. When changes are made to a dataset, it is useful to save the data with the date written into the file name. That way, it is possible to go back to previous iterations of the data to check for mistakes in the coding and recoding process. When files are merged, we make sure to check each variable by obtaining descriptive statistics, then run a list of the new or modified variables in a subsample of the data and compare this against the unmodified file in order to check the new dataset for accuracy. Occasionally a program file will become corrupted and unusable. For this reason, backing up command files/logs into a word-processing program is highly recommended. Keeping track of analyses and coding schemes that did *not* work – as well as the ones that did – will save time when working with the data in the future.

Step Three: Creating Needed Variables

Some secondary datasets were collected for purposes other than quantitative analyses, and contain qualitative information in non-numeric, text format. These types of qualitative data can be recoded using statistical software to assign numeric values to these variables. In our Civil War soldiers SD project, doing this correctly required a standardized coding scheme and permanently labeled new variables (i.e., all 'gunshot wounds to the arm' = 1; all 'head injury due to combat' = 2). Systematically coding and labeling our new numeric variables ensured that we could remember how or why a variable was created when looking back on analyses later.

There may be historical or economic events that influence the SD and should be accounted for in statistical analyses (e.g., when the data were recorded and what variables were collected). For example, in our longitudinal (1861–1910) Civil War project, the government military pension laws changed around the turn of the nineteenth century. The new law increased the types of pensionable medical conditions that a veteran could report in his medical record. We statistically accounted for this significant event by examining veterans who lived past the change in law so that all participants had the opportunity to report their full medical conditions.

Composite variables

Many SD sets include the individual items from larger complex measures. For example, the NLSY79 SD contains data on the individual items from a scale of depressive symptoms (CES-D) and the user must combine the 20 items to create a composite score. When doing this, it is important to note any items that are reverse coded and make sure to re-order the values of these items before creating the composite score. If the data repository does not provide the standard instructions on how to re-order the variables, that information can be found in the test designer's original documentation; this information is often available on-line or can be requested from the scale's publisher.

Another general type of composite variable, an 'index', can be created to represent a construct of interest via counting the occurrence of some characteristic or by the creation of a variable based on an individual meeting some threshold within the available data.

For example, employing SD, Dooley et al. (2005) used a count of the number and type of symptoms a person endorsed regarding alcohol use to create categories of alcohol dependence and abuse according to DSM-IV criteria. In another example, Kubzansky, Koenen, Jones, and Eaton (2009) examined the threshold effect of posttraumatic stress disorder (PTSD) symptoms on the development of cardiovascular disease. Three categories of symptoms were created (low, moderate, or high) and, based on the number of symptoms reported, placed the participants into one of the three groups.

Proxy variables

In working with our Civil War SD, we created a proxy variable to represent the primary construct of interest: traumatic stress. Because the literature in the area of traumatic stress has demonstrated that facing the threat of personal death and witnessing death during combat is objectively stressful, the number of recruits that were killed in a soldier's company during the war was used as a proxy for being threatened with and witnessing death. This variable (percentage of death in a soldier's company) was predictive of negative long-term health problems such as cardiovascular disease. Using the same SD, economists and historians have created proxies for the construct of social support by merging census and military records from the nineteenth century (i.e., the distance an individual lived from a group of similar ethnicity; how many extended family members lived within a close distance to the individual) (Costa & Kahn, 2008). This example demonstrates how psychological research can be enriched by exploring trans-disciplinary conceptualizations of constructs such as social support.

Step Four: Statistical Considerations

There are methodological considerations that arise more frequently when using SD but these issues are certainly not unique to its use. Among these considerations are the treatment of missing data, utilization of sampling weights, and being thoughtful about the statistical consequences of working with very large samples. Andersen, Wade, Possemato, and Ouimette (2010) used SD to examine the association between PTSD and physical disease in over 4,000 Iraq and Afghanistan war veterans. The large sample size was beneficial because it allowed for precise estimation of population values and provided good statistical power for planned comparisons (e.g., probability of detecting an association that exists in the population). However, an important consideration with large sample sizes is the risk of having too much power and of detecting very small, perhaps trivial, effects as statistically significant. Because the project was examining mental and physical disease conditions, this team considered the significant findings in terms of clinical relevance. They found that within the first 5 years of returning from war, veterans with PTSD were at over 30% increased risk of hypertensive and digestive disease conditions. Compared to a statistically significant but weak association (1–2% increased risk), an effect of this size is clinically meaningful for health care providers when considering physical disease prevention and treatment programs for veterans with PTSD.

To determine the practical magnitude of significant findings, it can also be helpful to compare the observed effect to standardized effect sizes (Cohen, 1988) or to a known effect size as presented in similar published studies. Importantly, even with large sample sizes, there may not be adequate power for subset analyses or to examine interactions between variables of interest. Although their large secondary dataset included both female and male veterans, Andersen et al. (2010) did not have adequate power to examine interactions between gender, PTSD and disease conditions among the smaller subsample of women.

Missing data

Secondary datasets, particularly in longitudinal studies, often have missing data. The SD repository should have information about the amount and location of missing data in the dataset. Although this is an unavoidable problem, many commonly used statistical software programs offer various options for handling missing data. One conventional approach deletes all observations with a missing value on any one of the variables being used in the analysis (i.e., listwise deletion, complete case analysis). This approach assumes that the missing data are 'missing completely at random' (MCAR), which means that the subset of subjects with complete data represent a random sample of the original set of observations (Alison, 2001). Unfortunately, this assumption is not often met and if missing data are scattered about many observations, this approach can reduce the sample size substantially, leading to inefficient use of the data and reduced statistical power. There are different types of missing data mechanisms (Rubin, 1976) and some statistical techniques make assumptions about the type of mechanism (e.g., Linear Mixed Models assume the data are 'missing at random' [MAR], a much weaker [less restrictive] assumption than MCAR) (Verbeke & Molenberghs, 2000; West, Welch, & Galecki, 2007).

In contrast to MCAR, under an assumed MAR missing data mechanism, the subset of cases with complete data is not assumed to be a random sample of the original set of observations. This distinction has implications for the validity of estimation procedures because maximum likelihood (ML) will produce valid parameter estimates if the missing data mechanism is MCAR or MAR, although for MAR it is necessary to assume that the specification of the joint distribution of the responses is correct. Under generalized least squares estimation (GLS), parameter estimates are valid for MCAR but can be biased for assumed MAR mechanisms. As a general rule, it is important to assess the amount and mechanism of missing data because this information will help inform the choice of a statistical method and the decision to use a missing value imputation procedure (Little & Rubin, 2002). We strongly recommend consulting with an individual well-versed in these statistical issues prior to beginning one's data analysis.

Imputing data

Procedures for 'imputing' values for missing data values have been developed that use information from observed data to create data points that are used to fill-in missing values. One of these imputation methods, mean substitution, substitutes the mean of the variable into the missing values for the variable. This method is generally not recommended because it ignores observed information about the subject for whom the mean is substituted and it can lead to biases in the standard errors. More rigorous multiple imputation procedures have been developed that use the original data to create several different datasets, each with the missing values imputed. The general idea behind these procedures is that observed data from a specified set of predictor variables are used to 'predict' or 'impute' values for the missing values. Although these procedures assume the missing data mechanism is MAR, they are often more desirable because they use observed subject-specific data to create imputed values. Multiple imputation programs are available in some of the commonly used statistical software, though many of these programs are based on different statistical assumptions (e.g., some assume a multivariate joint distribution while others do not; Royston, 2005). Often a form of 'sensitivity analysis' (analyses conducted under different assumptions) can be helpful in determining how vulnerable the results of the statistical analysis are to the imputed values or to the assumed missing data mechanism (Carpenter & Kenward, n.d.; Rosenbaum, 2002). Data workshops are often sponsored by academic institutions that profile various techniques for managing missing data when using SD.

Survey weights

Because SD often originate from surveys with complex sampling designs, 'weights' are frequently made available to users of the dataset. The variables storing the values of the weights for the cases in the dataset should be identified and downloaded along with the other variables needed for the research project. When used properly, these weights re-distribute the sample to be representative of a larger, well-defined population, which strengthens the external validity of the study findings. Misuse or failure to use the weights can result in serious biases to parameter estimates and can impair the generalizability of the results. Statistical analysis of survey data should account for the probability of inclusion and any clustering and/or stratification present in the data (Vittinghoff, Glidden, Shiboski, & McCulloch, 2005). Statistical software programs allow use of a variety of different weights while performing both descriptive and regression analyses. When using these types of SD, it is very important to become familiar with the types of weights available and how they are used. Workshops made available to users of the dataset can be very helpful as these often explain how to use the weights; networking about weighting issues with other researchers using the SD set can also be very productive. Statisticians or statistical consulting centers are another excellent resource for assistance with weighting issues.

A Caveat: Conflicting Messages from a Single SD

A concern when working with SD is the fact that there are sometimes different or conflicting results reported from research groups utilizing the same dataset. For example, Dohrenwend et al. (2006) conducted a remarkable reanalysis of data from the 1988 National Vietnam Veterans Readjustment Study (NVVRS), a landmark project, and the only fully representative sample of US veterans of the Vietnam War. The original study reported a 30.9% lifetime and 15.2% current rate of PTSD among veterans. Dohrenwend et al. (2006) developed an index of record-based war trauma exposure (e.g., military occupation, monthly and unit causality rate) and utilized clinician-diagnosed PTSD rather than examining self-reported trauma exposure and PTSD symptoms, as was done in the original study. Results showed 18.7% onset and 9.1% current rates of PTSD. This discrepancy sparked great controversy over the possible falsification or misrepresentation of early data findings. In this situation it is important to note that the methodologies and measures of both predictive and outcome variables varied between research groups. Despite differences, the direction of the findings were the same across studies. The reanalysis provided further evidence that Vietnam Veterans experience elevated rates of PTSD compared to civilians, and trauma characteristics play a substantial role in PTSD onset. These important similarities were clouded from the public by controversial debate about the differences in percentage rates of PTSD from one analysis of the dataset to the next.

When there are multiple publications from the same data source there are bound to be a variety of findings and interpretations, some of which can be directly opposite; this creates a challenging conundrum. While there is no magic solution to this problem, there are questions that can be asked in order to understand why there are contrasting findings. When discrepancies in the literature are encountered, it is important to compare how each group formed and operationalized their constructs of interest. What variables, indices, proxies, or scales did each group use to test their hypotheses? If the constructs and variables/scales are the same, check the reliability estimates of both groups' variables/scales. Did one of the groups use only a subset of the sample that differs from other research groups? How did each group treat missing data: did some perform imputations while others did not? What and how many analyses did each group run? Are there other published articles using the same data that find results in accordance with one of the studies? Often, when each study is scrutinized at this level, there are some aspects that differ dramatically and they are likely to contribute to the different findings.

At times there are even disagreements between researchers on the same project regarding what data are clinically or otherwise important and how the results should be interpreted. For example, the National Institute of Child Health and Human Development (NICHD) funded a large team of experts to conduct the largest and longest running study of early childcare and youth development in the U.S. This longitudinal study, spanning decades, examines the impact of family and day care environments on child development. There was great controversy within the research group regarding the interpretation of data on the positive and negative impact of daycare on child development. One researcher presented data to the popular press that supported a particular theoretical interpretation, while other team experts critiqued this presentation as preliminary and based on possibly inadequate statistical analyses or flawed interpretation of the analyses conducted. A comprehensive bibliography of study details and findings is available from the NICHD website if one is interested in following the debate. While this type of disagreement is unsettling, it is a real possibility when working with SD. As an academic or layperson contemplating scientific research findings, it is important to be aware that results are subject to the framing or interpretation of the researcher and sometimes even the popular press. Results may differ because of different conceptualizations of variables, use of different subsets of observations from the larger sample, the type of statistical techniques and subanalyses that are conducted, or possible differences in the treatment of missing data or imputed values. Awareness of this potential problem highlights the importance of keeping detailed and accurate records of how one conceptualized, recorded, analyzed and made inferences from the data. A benefit of public scrutiny over research findings is a reminder of the importance of scientific integrity and transparent methodology.

Conclusions

Utilizing SD requires meticulous documentation and the ability to manage large, often complex sets of data effectively. Despite these considerations, the benefits are numerous. SD introduces multi-disciplinary perspectives into psychological research, provides access to well-collected, longitudinal, and diverse samples, and makes available data that required significant resources and time to compile. Overall, the use of SD in conjunction with studies utilizing other methodologies provides an excellent opportunity to round out a program of research.

Short Biographies

Judith Pizarro Andersen received her doctorate in Psychology and Social Behavior from the University of California, Irvine. Currently she is a Post-Doctoral Associate at Cornell

University where she conducts research and teaches courses such as Health Psychology. In her research she examines the biological and psychological mechanisms by which the experience of extreme stress influences physical health across the life span. She has been funded by the University of Chicago's Center for Population Economics, National Institute of Aging grant to examine the mental and physical health effects of war experience on the life-span health of Civil War Veterans. In addition to longitudinal data analysis on population samples of adults exposed to the terrorist attacks of September 11^{th}, 2001, she has conducted experimental laboratory research on the biological mechanisms of stress in samples of women of childbearing age. She has published this work in journals such as the *Archives of General Psychiatry*, *Health Psychology* and abstracted in *Science*. Before coming to Cornell University, Andersen was a research scientist at the Department of Veterans Affairs Medical Center where she examined the physical health outcomes of veterans with Post-traumatic Stress Disorder.

JoAnn Prause received her doctorate in Social Ecology from the University of California, Irvine and has a Masters degree in Biostatistics and Quantitative Epidemiology. Prause is an expert in statistical and quantitative methods and teaches courses such as multiple regression, logistic regression, longitudinal data analysis, and hierarchical linear modeling at the University of California, Irvine. She maintains an extensive multidisciplinary research program that includes studying the effects of unemployment and inadequate employment and other stressors on mental health, alcohol misuse, and birth weight. She has authored and co-authored papers in journals such as *Psychological Bulletin*, *American Journal of Community Psychology*, and the *New England Journal of Medicine*. Before coming to UCI she was a Senior Biostatistician at Cedars-Sinai Medical Center in Los Angeles, California. There she provided statistical consultation in the areas of experimental design, statistical methods and analyses and interpretation of studies in clinical medicine.

Roxane Cohen Silver, Ph.D. is a Professor in the Department of Psychology and Social Behavior and the Department of Medicine at the University of California, Irvine. An international expert in the field of stress and coping, Dr. Silver is a Fellow of both the American Psychological Association and the Association for Psychological Science. At UC Irvine, Professor Silver has been actively involved in research, teaching, mentoring and administration. In recent years, she has become increasingly involved in and committed to bringing social and behavioral science theory and research into practice at the highest levels of government. In 2007, she received the American Psychological Association's Award for Distinguished Service to Psychological Science. Dr. Silver investigates the psychological, physical and social impact of traumatic life experiences, including personal losses as well as larger collective events such as natural disasters, terrorist attacks, and other community traumas. She was principal investigator of the only national longitudinal study of psychological responses to the September 11th terrorist attacks; the first report appeared in the *Journal of the American Medical Association (JAMA)* in September 2002. Her research has been funded by the National Science Foundation, the National Institute of Mental Health, and the U.S. Public Health Service (Bureau of Maternal and Child Health). Professor Silver completed her undergraduate and graduate training in Social Psychology at Northwestern University, Evanston, Illinois, and was on the faculty at the University of Waterloo, Ontario, Canada, before relocating to UC Irvine in 1989.

Endnote

* Correspondence address: Department of Psychology & Social Behavior, University of California, Irvine, 4201 Social & Behavioral Sciences Gateway, Irvine, CA 92697-7085, USA. Email: rsilver@uci.edu

References

Alison, P. (2001). *Missing Data*. Sage University Papers Series on Quantitative Application in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th edn). Washington, DC: American Psychiatric Association.

Andersen, J., Wade, M., Possemato, K., & Ouimette, P. (2010). Association between posttraumatic stress disorder and primary care provider-diagnosed disease among Iraq and Afghanistan veterans. *Psychosomatic Medicine*, **72**, 498–504.

Carpenter, J., & Kenward, M. (n.d.). Guidelines for handling missing data in social science research. Retrieved September 7, 2009, from http://www.missingdata.org.uk/.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Hillsdale, NJ: Lawrence Erlbaum.

Costa, D. L., & Kahn, M. E. (2008). *Heroes and Cowards: The Social Face of War*. Princeton, NJ: Princeton University Press.

Dohrenwend, P. P., Turner, J. B., Turse, N. A., Adams, B. G., Koenen, K. C., & Marshall, R. (2006). The psychological risks of Vietnam for U.S. veterans: A revisit with new data and methods. *Science*, **313**, 979–982.

Dooley, D., Prause, J., Ham-Rowbottom, K., & Emptage, N. (2005). Age of alcohol drinking onset: Precursors and the mediation of alcohol disorder. *Journal of Child & Adolescent Substance Abuse Research*, **15**, 19–37.

Feenberg, D., & Miron, J. (1997). Improving the accessibility of the NBER's historical data. *Journal of Business and Economic Statistics*, **15**, 293–299.

Freedland, K. E., & Carney, R. M. (1992). Data management and accountability in behavioral and biomedical research. *American Psychologist*, **47**(5), 640–645.

Holman, E. A., Silver, R. C., Poulin, M., Andersen, J., Gil-Rivas, V., & McIntosh, D. N. (2008). Terrorism, acute stress, and cardiovascular health: A 3-year national study following the September 11th attacks. *Archives of General Psychiatry*, **65**, 73–80.

Jordan, T. E. (1994). The arrow of time: Longitudinal study and its applications. *Genetic, Social, and General Psychology Monographs*, **120**, 469–531.

Kubzansky, L. D., Koenen, K. C., Jones, C., & Eaton, C. J. (2009). A prospective study of PTSD symptoms and coronary heart disease in women. *Health Psychology*, **28**, 125–130.

Little, R., & Rubin, D. (2002). *Statistical Analysis with Missing Data* (2nd edn). New York: Wiley-Interscience.

Pizarro, J., Silver, R. C., & Prause, J. (2006). Physical and mental health costs of traumatic war experiences among Civil War veterans. *Archives of General Psychiatry*, **63**, 193–200.

Rosenbaum, P. (2002). *Observational Studies* (2nd edn). New York: Springer-Verlag.

Royston, P. (2005). Multiple imputation of missing values: Update. *STATA Journal*, **5**, 188–201.

Rubin, D. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Tomlinson-Keasey, C. (1996). Opportunities and challenges posed by archival data sets. In D. C. Funder, R. D. Park, C. A. Tomlinson-Keasey & K. Widiman (Eds.), *Studying Lives Through Time, Personality, and Development* (pp. 65–92). Washington, DC: American Psychological Association.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Vittinghoff, E., Glidden, D., Shiboski, S., & McCulloch, C. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL: Chapman & Hall/CRC.

Wicke, T., & Silver, R. C. (2009). A community responds to collective trauma: An ecological analysis of the James Byrd murder in Jasper, Texas. *American Journal of Community Psychology*, **44**, 233–248.

Wilsnack, S. C., Hughes, T. L., Johnson, T. P., Bostwick, W. B., Szalacha, L. A., Benson, P., et al. (2008). Drinking and drinking-related problems among heterosexual and sexual minority women. *Journal of Studies on Alcohol and Drugs*, **69**, 129–139.

Wilsnack, R. W., Kristjanson, A. F., Wilsnack, S. C., & Crosby, R. D. (2006). Are U.S. women drinking less (or more)?: Historical and aging trends, 1981-2001. *Journal of Studies in Alcohol*, **67**, 341–348.

Wilsnack, R. W., Wilsnack, S. C., Kristijanson, A. F., & Harris, T. B. (1998). Ten-year prediction of women's drinking behavior in a nationally representative sample. *Women's Health*, **4**, 199–230.

Wilsnack, R. W., Wilsnack, S. C., Kristjanson, A. F., Vogeltanz-Holm, N. D., & Gmel, G. (2009). Gender and alcohol consumption: Patterns from the multinational GENACIS project. *Addiction*, **104**, 1487–1500.

Wolfe, W. R., Weiser, S. D., Leiter, K., Steward, W. T., Percy-de Korte, F., Phaladze, N., et al. (2008). The impact of universal access to antiretroviral therapy on HIV stigma in Botswana. *American Journal of Public Health*, **98**, 1865–1871.

World Health Organization (1999). *International Classification of Diseases, 9th Revision, Clinical Modification*, 5th edn. Los Angeles, CA: Practice Management Information Corp.