Data resources for high school sociology: Where to get data, and what students can do with it

Sean Kelly
University of Pittsburgh

Suggested Citation: Kelly, Sean. (2014). Data resources for high school sociology: Where to get data, and what students can do with it. A presentation at the University of Pittsburgh’s Sociology for Social Studies Fall Workshop. October, 21. Pittsburgh, PA.
Quantitative data analysis in high school social studies?

What used to be done in college is now being done in high school—Today’s students are learning data analysis skills at an earlier age

From the Common Core:

6th grade math standards have students posing and recognizing statistical questions about central tendency and variability, and graphing and describing distributions.

7th grade math standards have students taking random samples, informally comparing groups based on samples, and using basic probability concepts to understand the likelihood of events/chance occurrences.

By high school, students are expected to be making inferences from experimental and observational data.
Data analysis in high school social studies?

Yet, there are still some big hurdles to overcome:

- **Familiarity.** “Where did this data come from? All I see in this data is a variable name, that’s not much to go on. And is this even good data to begin with?”

- **Size.** “there are hundreds of variables here, how do I even know where to start?”

- **Software.** “I understand what we are studying, but this statistical software is overwhelming.”
The challenges of familiarity and size can be overcome by having students collect their own data! There are many other benefits to this approach as well:

- Students are in the driver’s seat for the whole research process, from posing questions, to identifying participants, to developing measures. This solves the familiarity problem, allows you to teach a broad array of skills, and gives students **ownership** of the data.

- Students have the opportunity to showcase a broader array of skills than just statistical analysis.

- If students are pooling data together, then the overall quality of the project demands a team effort.

In all, *Collecting your own data is fun and engaging!*
But there may still be a time to have students work with secondary data

- Secondary data gives students access to populations that they might not have access to in their community.

- Secondary data allows students to carry out historical/time-trend analyses.

- Secondary data poses all sorts of logistical challenges, but for students who aren’t completely stymied by those challenges, overcoming them can be a learning experience in itself.

- Secondary data can have caché: You get to work with the same data the “pros” use.

- If students find and select their own data/research questions (assuming that is appropriate), there is still an element of autonomy and creativity involved in that.

- Use of secondary data may be an expedient way to get students to engage with data terms (e.g. dependent variable) and concepts (error variance); students collecting their own data may be so in to the process that they don’t take the time to develop that vocabulary.
Using Secondary Data

What’s out there? Where can I find data and which specific social science databases should I consider using?

How should I use it? How much scaffolding is required? What are my time/effort costs up front?

Working with raw data vs. embedded data analysis systems.

If I decide to use raw data, what statistical analysis software should I use: Tradeoffs between Excel and SPSS.
Where Can I Find Secondary Data?

Major Social Science Surveys include:

- The General Social Survey (GSS): a very wide general topic survey including rich data on a variety of social attitudes.
- The American National Election Studies: a set of studies on voting and political behavior.
- The Behavioral Risk Factor Surveillance System (BRRFSS) & the National Health Interview Survey: two studies of health and well-being.
- National Survey on Drug Use & Health (formerly the NHSDA): A substance abuse study.
Where Can I Find Secondary Data?

Two government agencies that collect a wide-array of topical surveys:

- The Bureau of Justice Statistics
- The National Center for Education Statistics

In addition, there is the:

Inter-university Consortium for Political and Social Research (ICPSR): A clearinghouse of smaller, stand-alone surveys.

ICPSR has hundreds of studies archived. It can be a good source for finding urban/community studies of particular areas, and there are some real gems in there like the Quality of American Life Studies from 1971 and 1978. But anyone can upload a database to ICPSR, so the quality and usability is uneven at best.
The GSS

If there is one all-purpose, go-to survey for sociological inquiry in the U.S., it is the General Social Survey.

http://www.norc.org/GSS+Website/

The GSS has been conducted annually or every other year since 1972. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events. The GSS is an excellent and often-used source for sociological and attitudinal trend data covering the United States. There have been over 14,000 research articles using the GSS.
The GSS

The GSS is a household, face-to-face survey conducted on computer. It is nationally representative of the household population of the U.S. and is a probabilistic sample. Census tracts (neighborhoods) are randomly selected and each dwelling unit (house or apartment) is counted and listed. Units are selected and individuals 18 years of age and older randomly sampled within households. Until 2006, when a Spanish survey was added, only English speaking respondents were interviewed. The response rate averages about 70%, which is typical of household surveys.

GSS surveys consist of a core group of questions (including all the basic demographic and sociology type questions, family background, etc.) and modules of questions that are added by researchers to address other issues. These modules address very diverse topics.

All in all there have been over 57,000 respondents in the 30+ years of the GSS and there are over 5,000 variables. The non-core modules rotate on and off. This approach expands the number of items that have been included on the GSS, but it also interrupts the time-series for certain questions (not all questions are available in any given year).
You can browse the content of the GSS by subject (Alienation, Birth Control, Cohabitation, etc.) or by the topical modules that were used in a given year: “quality of working life” (2002, 2006) “altruism” (2002, 2004) or simply by year, and look at descriptive statistics for specific questions on the survey.

You can search for reports/articles, etc., written with GSS data too:

http://www.norc.org/GSS+Website/Publications/Advanced+Search/

**Documentation**

Full documentation is available at:
http://www.norc.org/GSS+Website/Codebook/

Changes to any questions over-time are documented in:  http://publicdata.norc.org:41000/gss/Documents/Codebook/N.pdf

How should students work with secondary data?

A big decision in working with secondary data is whether to work with raw data or to work with an embedded data analysis system.

Embedded data analysis systems allow students to work entirely in an on-line platform, exploring, querying the data, and producing results all in one place. Many data centers now have data analysis systems that can be used to access the data.

The General Social Survey as well as census and other survey data is supported by the Survey Documentation and Analysis (SDA) online system hosted by the University of California-Berkeley.

http://sda.berkeley.edu/archive.htm

Textbooks have been written to support analysis with the SDA, such as Feigelman and Young’s (2006) Hands on Sociology, 3rd edition, published by Pearson.
Here I’ve generated a quick chart that shows beliefs about the role of low quality schools (are poor schools an important source of poverty in America) among owners and renters. Those who own their own home appear to be less likely to cite low quality schooling as a major cause of poverty than those who rent.

When I say quick, I mean quick, you can make a chart like this in like 2 minutes!
Drawbacks to using an embedded data analysis system

Embedded data analysis systems have the clear advantage of accessibility, but they are not without drawbacks:

- It is possible to use a system like the SDA without spending any time exploring the data. In the previous example, the question about the role of schools in generating poverty was only asked of approximately 800 respondents in a single module, something a student might not even realize.

- Most SDAs have some analytic flexibility built in, but they are only so flexible; the student may not be able to combine variables, restrict analyses to specific respondents, etc., in the manner they may want to.

- When a student learns to use an SDA, they learn only to use…that specific SDA, not a software system they might use of a wide array of purposes for years to come, or general principles of database management they may need in their future careers.
But is it even possible to work with raw data?

Yes, certainly. But depending on the database, you may have to do some work up-front to prepare/clean the data for students to work with (or recruit a colleague to help you prepare data for joint use across the curriculum). Issues to consider:

- Try to prepare a database(s) with no more than a few hundred variables, ideally fewer.

- Make sure that data is stored in a useable format (e.g. numeric data is stored as numeric, not string [text] data).

- Consider recoding missing data codes (-9, etc.) to missing (e.g. “.”).

- Make sure data is fully labeled (variable name, variable label, and value labels).

- Distill documentation for your working database from the original documentation.
Here I will discuss two alternatives, **Microsoft Excel** and **SPSS**. Of course there are dozens of other possibilities, but these are the only real ones.

Excel needs to be seriously considered, because it is installed on virtually every computer in the world, and your students can be expected to have “free” access to it if they work with a computer.

Of the dedicated statistical analysis software packages, SPSS is far and away the most widespread, and it is a menu-driven, user-friendly program. If students have access to SPSS, it is certainly an option, and they can keep using it in college and beyond.
Both programs:

- Allow the user to “see the data” in rows and columns.
- Can perform all the basic statistical functions your students would need to analyze the data.
- Have “graphical design” capabilities; they can produce tables and figures.

Excel however is not a “database” program per se, it is a spreadsheet. The data is stored and queried by it’s location in a particular cell rather than as values of specific variables (thus in excel for example, you can calculate an average in a row, column, or even a block of rows and columns). Thus, it is:

- Difficult or cumbersome to perform some database management tasks
- Cumbersome and non-intuitive to perform some analyses.
- Difficult to work with large numbers of cases or variables.
In my opinion, whatever start-up costs are entailed in learning SPSS (or similar program), are worth it because learning the program reinforces the fundamental logic of statistical analyses.

But if a license to SPSS or other dedicated statistical software is not available, students can still engage in secondary data analysis in Excel, and you may just want to plan accordingly (e.g. using a smaller database rather than a larger one).

Indeed, Excel would work great for having students collect and analyze data as a class or small-groups.