Broadening Public Access to Data Through the Development of Tools for Data Novices

Josefina J. Card
Lauren Shapiro
Angela Amarillas
Elizabeth McKean
Tamara Kuhn

Reprinted from Social Science Computer Review
Volume 21, No. 3, Fall 2003, pp. 352-359

Broadening Public Access to Data Through the Development of Tools for Data Novices

JOSEFINA J. CARD
LAUREN SHAPIRO
ANGELA AMARILLAS
ELIZABETH McKEAN
TAMARA KUHN
Sociometrics Corporation

In this article, the authors describe three new resources aimed at bringing scientific data, data analysis, and data interpretation to nonexperts (e.g., high school students; undergraduates; health practitioners) through innovative web-based and CD-ROM–based tools. In the first resource, original survey questions and answers are used as the basis for searching through data collections instead of analytic software-generated variables and response codes. Search results are presented as data tables, allowing novices to access survey findings without having to conduct data analysis themselves. A second set of tools is aimed at reaching an even larger audience of nonexperts through the development of a research-based data and Internet literacy curriculum, along with associated training materials. Finally, a third approach is looking at bringing math concepts and statistical data to high school students through the use of real-world *DataStories* that come to life with the help of computer-based story scripts, animation, audio, and interactive exercises.

Keywords: health data; survey research; retrieval software; statistics instruction; Social Science Electronic Data Library

A leading barrier to secondary data use has been the difficulty of obtaining and understanding the documentation required to use the data set—assuming that such documentation exists, which is often not the case. A second challenge has been the difficulty of searching through available data sets to find studies and variables of use for a particular research problem of interest. Over the past 19 years, Sociometrics Corporation has tried to address these challenges by assembling the Social Science Electronic Data Library (SSEDL) with funding from various National Institutes of Health and from the National Science Foundation. Now available in CD-ROM and web formats, SSEDL includes more than 250 data sets from exemplary studies in nine health and social science fields: adolescent pregnancy and fertility, the American family, social gerontology, maternal drug abuse, AIDS and sexually transmitted infections, disability, contextual influences on behavior, child welfare and poverty, and complementary and alternative medicine. Scientist expert panels have selected the leading data sets in each of these nine fields, based on scientific merit, substantive utility for secondary data analysis, and program and policy relevance (Card, 1996, 2000, 2001; Card & McKean, 1995; Card & Peterson, 1991; Carley & Card, 2000).

Social Science Computer Review, Vol. 21 No. 3, Fall 2003 352-359
DOI: 10.1177/0894439303253983

SSEDL has developed and continues to improve a complete, standardized, and user-friendly approach to documentation as well as information search and retrieval. Several features are incorporated into the data sets comprising SSEDL to enhance their quick and accurate use for research.

Indexing at the variable level. Each variable in each data set is indexed according to a set of approximately 60 archive-relevant topics that characterize the substance of the variable (e.g., "marriage and divorce," "contraception") and approximately 15 types that characterize the kind of measure (e.g., "attitude," "behavior," "status"). The topics and types vary somewhat according to the particular topically focused collection in which a particular data set is placed. This topic-and-type classification affords users a powerful method of quickly searching for, and then extracting, variables of interest both within and across data sets in a collection.

Quality documentation. Each data set is made publicly available with a standard set of five machine-readable data and documentation files: File 1 is a raw data file; Files 2 and 3 are machine-readable SPSS and SAS program statements that fully document the variables and values in the data file;[1] File 4 is an SPSS data dictionary; and File 5 is SPSS frequencies. Each data set also is accompanied by a printed user's guide (also provided in machine-readable form) comprising a standard set of sections and subsections describing the study's goals, methods, and content. The provision of standard machine-readable and printed documentation assists users in familiarizing themselves with SSEDL data sets. Once a user has worked with one SSEDL data set, it is easy for him or her to work with any of the others. The original instrument and codebook are offered as optional supplementary documentation for each data set, when available. For all new data sets, the original instrument is distributed in machine-readable form along with the data, as a set of graphics files (page images).

Search and retrieval software. Web users can search the contents of SSEDL by keyword, free of charge, at the study level and at the variable level (`http://www.socio.com/search.htm`). Even more powerful search and retrieval software is found in the CD-ROM version of SSEDL. This software allows a user to search an entire topically focused archive, a group of data sets created explicitly for a given user, or a single data set; to identify variables of interest across this designated search space; and to save located variables as a search set. Users can conduct (a) full-text keyword searches, including variable names, words in variable labels (question descriptors), and words in value labels (response descriptors); (b) searches by assigned topic and type codes; and (c) searches by study name or assigned data set number. Standard Boolean operators (i.e., *and*, *or*, *not*) can be used to combine search sets.

These standard features make the data sets in SSEDL easy to use, even by novice analysts. Still, many students and professionals whose schoolwork or jobs could be enhanced by use of appropriate data are daunted by having to conduct data analysis with statistical analysis packages such as SPSS or SAS. In this article, we describe our recent work making SSEDL data sets even easier to use by these data novices through the development of innovative tools that facilitate data access, search, presentation, and interpretation.

FACILITATING THE INFORMATION ACCESS, SEARCH, AND PRESENTATION PROCESS

In SSEDL, as previously described, each variable is indexed or coded by its topic as well as its type. These topic and type codes, along with SPSS variable and value labels, are used to search for variables meeting user-specified search criteria. The search returns a list of variables that can be used to create a data extract file for analysis with SPSS. For our first resource for data novices, we changed the variable search parameters to simplify the information search even further, simultaneously enhancing the scope of the search.

Rather than searching by keyword or topic and type codes through SPSS variable and value labels, original survey questions and answers (in their original English format as seen by survey respondents) are used as the basis for browsing and searching through data collections. The user clicks on a question of interest and search results are presented as data tables rather than variable lists requiring further processing by an analytic package such as SPSS or SAS. These two innovations—using original survey questions as the basis of the search and data tables as the output of the search—allow novice users to access survey findings without having to conduct data analysis themselves.

This resource is contained in a book with a CD-ROM titled *U.S. Social Surveys: Questions and Responses From National Studies* (McKean et al., 2003). The resource links data collection questionnaires and data tables from the six best-selling national studies in SSEDL: National Health and Social Life Survey, 1992; National Longitudinal Survey of Adolescent Health, 1994; National Survey of Family Growth, Cycle 5, 1995; National Survey of Adolescent Males, 1995; and Youth Risk Behavior Surveys, 1997 and 1999. Users can peruse the questionnaire from any of these six studies, click on questions of interest, and view, print, or download the following: (a) a study summary that describes how the study was conducted; (b) bar charts displaying response distributions for the selected questionnaire item; (c) data tables displaying the cross-tabulation of responses by age, race, and gender subgroups; (d) a teaching module comprising a series of progressively more difficult exercises that illustrate various statistical and social science methods or concepts; (e) a machine-readable data extract file with SPSS syntax to recreate the teaching module exercises or perform new data analyses; (f) frequently answered questions about the data set; and (g) a glossary of useful terms.

DEVELOPING A DATA AND INTERNET LITERACY TRAINING CURRICULUM FOR THE DATA NOVICE

A second resource for data novices built around SSEDL was developed to reach an even larger audience of data novices. The *Data and Internet Literacy Series* is a research-based syllabus and curriculum with accompanying educational materials (lecture slides, lessons, student manuals, instructor packages) to help make data novices data and Internet literate at an appropriate minimum level required by their educational training program or by their jobs. The training materials have been made available for distance learners on the Internet and also are available as supplementary course materials for use in undergraduate classrooms. When developing this resource, we relied on results of needs assessment research to investigate the data and Internet training needs of data novices both inside and outside academia. The needs assessment procedure involved a six-step process.

1. We identified a range of populations where individuals are in regular contact with social science data at work. These populations included *university faculty*, including faculty within the disciplines of (a) sociology, (b) education, (c) psychology, (d) population/demog-

raphy, (e) social work, (f) public health, and (g) behavioral medicine; *university students* in the seven disciplines specified above, including (a) undergraduates, (b) master's students, and (c) doctoral students; *librarians*, including librarians at (a) universities and (b) public libraries; and *professionals in the field of behavioral public health*, including (a) philanthropic foundation staff, (b) governmental agency staff, and (c) intervention agency staff.

2. We selected core needs assessment participants through a variety of procedures. *Core faculty*: Six universities were selected randomly from among 77 SSEDL-subscribing institutions. Participants were chosen randomly from among those faculty whose names and e-mail addresses could be located through the web sites of these six universities. Insofar as these schools provided contact information for faculty members in all seven of the disciplines listed previously, participants were drawn from each discipline at each school. In addition, all faculty members who recently had placed an order for SSEDL (on behalf of any institution) were chosen. *Core students*: Participants were chosen randomly through their university web sites, according to the same procedures used for faculty. In addition, some faculty members were asked to nominate student participants. *Core librarians*: Participants were chosen randomly from among those librarians whose names and e-mail addresses were listed on the web site of the American Library Association. In addition, all librarians who recently had placed an order for SSEDL (on behalf of any institution) were chosen. *Core public health professionals*: Participants were chosen from among the staff of private and governmental organizations with which Sociometrics has had professional contact in the past.

3. One representative of each needs assessment subpopulation, except students, was contacted for a 20-minute telephone interview aimed at generating a list of basic data- and Internet-related concepts and skills of importance to academics and other professionals in the social sciences. In total, 12 interviews were conducted,[2] using a semistandardized interview script.[3] One version of the script was tailored to suit academic faculty, while a second version was tailored to suit nonacademic professionals (i.e., librarians and public health professionals). These interviews yielded a long list of data- and Internet-related concepts and skills that academics and professionals in the social sciences found important, as well as information about the instructional formats that they tend to favor.

4. The concepts and skills generated through telephone interviews were organized into content areas, including (a) data comprehension, (b) data collection, (c) data analysis, (d) statistics, (e) data interpretation, (f) data presentation, (g) primary and secondary data sets, and (h) the Internet. An online survey aimed at prioritizing the concepts and skills was developed based on these content areas, with questions framed slightly different for faculty, students, and professionals:[4] The faculty questionnaire asked, "How important is it for undergraduates who take classes within your department to achieve some level of competence in the following [specific content area] skills or concepts?" Faculty were asked to rate each item as *extremely important*, *somewhat important*, or *not important*. The student questionnaire asked, "How proficient would you judge yourself to be in the following [specific content area] skills or concepts?" Students were asked to rate themselves on each item as *extremely proficient*, *somewhat proficient*, or *not proficient*. The professional questionnaire asked, "How useful do you find the following [specific content area] skills or concepts in the course of carrying out your work?" Professionals were asked to rate each item as *extremely useful*, *somewhat useful*, or *not useful*.[5] In addition, the survey asked how comfortable participants would feel engaging in (or asking students to engage in) various activities as part of a course or tutorial. Overall, each version of the survey contained 87 dependent variables. The three questionnaires were posted on the Internet (each at a distinct web address) so that they could be accessed at any time from anywhere across the country.

5. Ninety-one faculty, 37 students, and 43 professionals were contacted by e-mail and invited to participate in the appropriate version of the online survey, in addition to 31 SSEDL subscribers (a mix of university faculty and librarians). In appreciation for their time, faculty and professionals were offered a $5 gift certificate to Amazon.com and a coupon for 50% off the cost of the training modules that result from the survey; students were offered only the gift certificate. The questionnaires remained online from December 6 to December 31, 2001. Sixty individuals responded, including 21 faculty, 24 students, and 15 professionals, for an overall response rate of 27.2%. The rate is in line with other rates reported for web-based surveys (Sills & Song, 2002), despite the fact that the survey was administered at the end of the academic term, during the winter holiday season.

6. The aim of data analysis was to identify concepts and skills that faculty found important and professionals found useful, for which novices (i.e., students) also showed low proficiency. To this end, weightings were assigned to each response category, and a mean was generated for each of the 87 variables within each of the three samples (for a grand total of 261 means). Specifically, *extremely* was weighted as 1, *somewhat* was weighted as 2, and *not* was weighted as 3. Particular attention was paid to items that achieved a mean rating of less than 2.00 for both the faculty and the professional samples.[6] Because preliminary analysis revealed that the student sample consisted primarily of doctoral students rather than data novices, less attention was paid to student ratings than originally intended. The concepts and skills that achieved the criterion of less than 2.00 were organized into a syllabus comprising 10 content areas (see appendix).

From this syllabus of concepts, a curriculum of six teaching modules containing these concepts was then developed: *Making Sense of Scientific Articles* (Shapiro, 2003b), *All About Data Collection* (Shapiro, 2003a), *Understanding Data in Numbers, Words, and Pictures* (Amarillas & Mince, 2003), *Using the Internet to Find the Information You Need* (Bunch, 2003), *How to Give Your Organization an On-line Presence* (Kuhn, 2003), and *Using Your Internet and Data Skills to Achieve Fundraising Goals* (Dull Akers, 2003). We also developed an instructor package for each module, consisting of a lesson plan, a set of PowerPoint slides, review questions and answers, and photocopy masters of the module's activities and review questions, to facilitate use of the module in a classroom setting.

USING DATA STORIES TO BRING HIGH-LEVEL MATH CONCEPTS TO HIGH SCHOOL STUDENTS

With funding from the U.S. Department of Education, we are conducting a feasibility study for a third resource called *DataStories*, aimed at enhancing math instruction and learning among high school students in the career and technical education (CTE) track. Our completed *DataStories* prototype consists of three parts. Part 1, Research Findings, is an electronic encyclopedia composed of multimedia presentations of salient up-to-date facts about adolescent health. Part 2, Research Data, contains extracts of national studies in SSEDL and a multimedia presentation of the goals and content of the included data files. Part 3, From Data to Findings, the heart of *DataStories*, is a multimedia, hands-on, interactive statistics tutor. Using problem sets of gradually increasing difficulty, the tutor teaches students how to do elementary analyses of the data in Part 2 to arrive at conclusions similar to those presented in Part 1. Once fully developed, *DataStories* will contain a set of self-contained Internet and CD-ROM–based teaching modules, each focused on one key concept or procedure included in math courses taken by CTE students. The hope is that *DataStories* will improve instruction and achievement in high level math with its combination of state-of-the-art technology, real-world data, research questions of interest to adolescents, stories, scripts, pictures, graphs, and interactive pedagogical techniques. A controlled field test study of the prototype module on correlation is currently being conducted at two high schools in California.

CONCLUSION

Increasing the science-based capacity of the nation is a priority of the National Science Foundation, the National Institutes of Health, and the Department of Education. By bridging the heretofore disconnection between the worlds of teaching, research, and practice, the tools described in this article move us closer toward achievement of this goal.

APPENDIX
Syllabus of Concepts and Skills Rated as Most Useful for Data Novices

I. *Making sense of scientific journal articles.* Understanding scientific journal articles, keeping up with current developments in the field, assessing the quality of a study, and identifying weaknesses in study design.

II. *Collecting your own data.* Applying published findings to one's own projects, understanding the logic of experimentation, designing an entire data collection project, determining which factors are relevant to measure, formulating survey questions, collecting closed-ended questionnaire data (e.g., multiple choice questions), collecting open-ended questionnaire data (e.g., essay questions), and collecting open-ended interview data.

III. *Thinking quantitatively: Key concepts in data analysis.* Translating qualitative data into numerical format, organizing numerical data into spreadsheets, organizing numerical data into cross-tabulation tables, comparing people in Group A with people in Group B, comparing pretest responses with posttest responses, determining whether two variables are related, understanding summaries of numerical data, and assessing the validity or adequacy of collected data.

IV. *What the numbers tell you: Analyzing data statistically.* Selecting the most appropriate technique for data analysis, mean/average, standard deviation/variability, effect size/statistical significance, statistical comparison of means, chi-square analysis, statistical analysis of correlation/associations between variables, using statistical software like SPSS or SAS, creating a codebook, and dealing with missing values.

V. *I've analyzed my data . . . but what does it mean?* Supporting or disproving a hypothesis, determining whether an intervention made progress toward pre-established goals, determining whether an intervention program has affected a community, developing a comprehensive picture of a problem or issue, making policy recommendations, and identifying important directions to pursue in the future.

VI. *How to present data persuasively.* Integrating numbers with qualitative descriptions; selectively using numbers to make a point; displaying data summaries (e.g., percentages); creating graphs, charts, or tables; including relevant data in a report or article; and using PowerPoint software.

VII. *Finding high-quality information online.* Accessing journal articles through online databases; locating answers to questions that arise during the day; locating information to support the writing of reports, articles, or proposals; locating recently updated statistical information; evaluating the quality of information found on a web site; evaluating the credibility of web site authors and sources; and compiling a list of the trustworthy web sites that exist.

VIII. *Using secondary data.* Locating useful secondary data sets, determining which secondary data sets are relevant to one's needs, making use of the secondary data sets that are available, using a data set codebook, using secondary data to situate a project in a larger context, and incorporating secondary data into reports or proposals.

IX. *Using your data and Internet skills to achieve fund-raising goals.* Understanding the data presented in grant applications, understanding financial statements and budgets, using data to argue that a person or team deserves financial support, including relevant data in proposals for financial support, using the Internet to locate funding opportunities, and using the Internet to identify content areas for future work.

X. *How to present yourself online.* Creating a basic web page.

NOTES

1. As an alternative to Files 1, 2, and 3, Sociometrics also will provide an SPSS portable file, which contains the same information and can be imported into SAS and a number of other data analysis programs.

2. No behavioral medicine faculty member was available by telephone, but a second education faculty member was interviewed.

3. Copies of the interview scripts are available upon request.

4. Copies of the questionnaires are available upon request.

5. With regard to statistics only, professionals also were given the option of *not sure*.

6. Information on mean scores is available upon request.

REFERENCES

Amarillas, A., & Mince, J. (2003). *Understanding data in numbers, words, and pictures*. Los Altos, CA: Sociometrics Corporation.

Bunch, M. (2003). *Using the Internet to find the information you need*. Los Altos, CA: Sociometrics Corporation.

Card, J. J. (1996). Development of the Sociometrics data library on families, aging, substance abuse, and AIDS. *Social Science Computer Review, 14*, 305-309.

Card, J. J. (2000). Development and dissemination of an electronic library of exemplary social science data. *Social Science Computer Review, 18*, 82-86.

Card, J. J. (2001). Archiving: Ethical aspects. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 646-649). Amsterdam, the Netherlands: Elsevier.

Card, J. J., & McKean, E. (1995). Development of software to facilitate use of archived data sets. *IASSIST Quarterly, 19*, 4-11.

Card, J. J., & Peterson, J. L. (1991). Establishing and operating a social science data archive. In J. E. Sieber (Ed.), *Data sharing: Advantages and challenges*. Newbury Park, CA: Sage.

Carley, M., & Card, J. J. (2000). The Social Science Electronic Data Library: Serving the needs of data librarians and users. *IASSIST Quarterly, 24*, 8-14.

Dull Akers, D. (2003). *Using your Internet and data skills to achieve fundraising goals*. Los Altos, CA: Sociometrics Corporation.

Kuhn, T. (2003). *How to give your organization an on-line presence*. Los Altos, CA: Sociometrics Corporation.

McKean, E. A., Card, J. J., Dull Akers, D., Thyer, B., Espinoza, R., Benner, T., & Peterson, J. L. (2003). *U.S. social surveys: Questions and responses from national studies*. Los Altos, CA: Sociometrics Corporation.

Shapiro, L. J. (2003a). *All about data collection*. Los Altos, CA: Sociometrics Corporation.

Shapiro, L. J. (2003b). *Making sense of scientific articles*. Los Altos, CA: Sociometrics Corporation.

Sills, S. J., & Song, C. (2002). Innovations in survey research: An application of web-based surveys. *Social Science Computer Review, 20*, 22-30.

Josefina J. Card, Ph.D., is founder and president of Sociometrics Corporation. E-mail: jjcard@socio.com.

Lauren Shapiro, Ph.D., recently received her Ph.D. in psychology from Stanford University and is currently a senior research associate at Sociometrics Corporation. E-mail: lauren@socio.com.

Angela Amarillas, M.A., is a research associate at Sociometrics Corporation, and her work has focused on the quest to bring social science research and data to novices, including promoting the Social Science Electronic Data Library, writing and editing for the Data and Internet Literacy Series, and developing Data Stories for high school math students. E-mail: angela@socio.com.

Elizabeth McKean, M.A., data division director at Sociometrics Corporation, also serves as Sociometrics' data archiving manager and director of the company's research support group. E-mail: `lmckean@socio.com`.

Tamara Kuhn, M.A., director of Internet development at Sociometrics Corporation, is an experienced social scientist and web developer who specializes in the design and programming of social science–related technology. E-mail: `tkuhn@socio.com`.