Journal of Official Statistics, Vol. 18, No. 1, 2002, pp. 111–123

Book and Software Reviews

Books for review are to be sent to the Book Review Editor Jaki Stanley McCarthy, USDA/NASS, Research Division, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A.

Theodore Caplow, Louis Hicks, and Ben J. Wattenberg. *The First Measured Century: An Illustrated Guide to Trends in America, 1900-2000*. Washington, D.C.; The American Enterprise Institute Press, 2001. ISBN 0-8447-4137-X, 20 USD.

This book is a ''companion volume to the three-hour PBS television documentary ''The First Measured Century,'' a prime-time special by the producers of the PBS discussion series ''*Think Tank*.'' It is a reference book that addresses the trends for the century for the subject areas of population, work, education, family, living arrangements, religion, active leisure, health, money, politics, government, crime, transportation, business, and communications. The sources of material for this volume are government tables and charts and the results of the *Middletown surveys* conducted in the 1924 with replicates in 1977 and 1999. These surveys were random samples of married couples with children under age eighteen in Munice, Indiana. The results were used to provide information for America and supposed trends for the century for topics not covered by the official statistics, such as, time mother and father spent with children in 1924, 1977, and 1999.

The twentieth century is the first century that was measured. This book presents the trends and social change that took place in the twentieth century. It is comprised of a Preface, fifteen chapters that address different subject areas of change during the century, Notes which present the background documents for these statements, an Index, info about the Authors and a section on Supplementary Resources. ''Each chapter is a topic area and within each chapter are a series of key trends, each explained in a one-page essay and illustrated by one or more colored graphs or charts on the facing page.'' Some information presented in the essays is not presented in the graphs or charts but the reference source is presented in the Notes. Some graphs would be better representatives of the time frame if they had been histograms. Each chapter could stand alone and addresses

the changes in the twentieth century for the subject area. The authors state that this book is a stand alone document and is not just a supplement to the PBS TV program. This book is clearly written and does a good job of presenting information on each subject area. It is fairly comprehensive in addressing each subject area, although it is not clear to me why the order of the subject areas is as it is. It does not require any special training to read and understand and, although it mainly presents ''facts and figures,'' does present plausible explanations for the changes over time. It would be of interest to anyone who has an interest in the history of the twentieth century, whether as research for some sort of paper or article or just to read. Overall, this book would make a ''good reference for teachers or students, journalists and bureaucrats, social scientists and others'' who want an overview of trends in the twentieth century or background information for the twenty-first century.

Judith M. Conn
Office of Statistics and Programming
National Center for Injury Prevention and Control
Centers for Disease Control and Prevention
4770 Buford Highway
Atlanta, GA 30341
U.S.A.
Phone: 770-488-4752
Fax: 770-488-1665
e-mail: jmc1@cdc.gov

Jeffrey A. Gliner and George A. Morgan. *Research Methods in Applied Settings: An Integrated Approach to Design and Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers (2000), 439pp + refs and index. ISBN 08058-2992-x, 49.95 USD.

Gliner and Morgan have produced an applied research methods text suitable for an upper-division or first-year graduate course in the applied behavioral sciences. The authors have varied experience teaching statistics and research design in several subject areas, and have designed the text to be applicable to courses in different departments such as psychology, social work, consumer science, and occupational therapy.

This book consists of 24 chapters divided into six major units: I. Introductory Chapters; II. Research Approaches and Designs; III. Understanding the Selection and Use of Statistics; IV. Integrating Designs and Analyses: Interpreting Results; V. Measurement, Instruments, and Procedures; and VI. Research Validity, Replication, and Review. Each unit contains several related chapters that encompass the full range of research designs and associated statistical procedures.

A positive feature of this text is its very clear and coherent organization. Each chapter begins with an outline of the major headings, minor headings, and sub-sections and concludes with a ''Study Aids'' section including a list of concepts, a list of important distinctions, and a number of application problems, emphasizing definitional, conceptual, and computational problems. The text is written for easy consumption and each chapter's content is well laid out with easy-to-follow sub-headings. In addition, the authors have supplied numerous applied examples from diverse fields, as well as exercises that cover several content domains.

The authors state that a primary goal of their volume is its ''student-friendly'' focus, and they have designed the text with a significant pedagogical emphasis on student comprehension of important research concepts. Acknowledging that many students do not see the link between a study's research design and the concomitant selection of an accompanying statistical procedure, they have attempted to draw this important conceptual link with the intention of making this relationship clearer to students. In an effort to simplify students' comprehension of complex topics, consistent terminology – what they call ''semantically consistent'' language – is used to describe research approaches (experimental, comparative, etc.), research questions (difference, associational, and descriptive), and types of statistics stemming from the application of these research methods and questions.

Another feature of the book is the categorization of the major research procedures into one of five categories that they call ''research approaches:'' 1) randomized experimental; 2) quasi-experimental; 3) comparative; 4) associational; and 5) descriptive. They acknowledge that complex studies may fall into more than one of these research approaches, and provide exercises that exemplify such complex studies and selection of methodologies.

The authors' primary focus is to help students become good consumers of research with an emphasis on the analysis and evaluation of the results of research articles. To this end, the book is designed to be *read*, rather than to be a purveyor of extensive exercise sets requiring the student to practice complex statistical analyses. Exercise sets for each chapter include a variety of conceptual and quantitative exercises designed to promote students' ability to become good consumers of research. This is an important distinction and should be considered when weighing the utility of this text for a given course. If you are seeking a book that will help students in applied fields understand how research is designed and results are analyzed, emphasizing students' ability to *consume* and *evaluate* research, this text is likely to be a good choice. A breakdown of each chapter's contents follows.

Chapter 1 provides an overview of the definitions, purposes, and dimensions of research, and serves to introduce the student to the material covered within the text. Chapter 2 does an excellent job of introducing both quantitative and qualitative research paradigms, emphasizing philosophical distinctions between both paradigms. Chapter 3 presents ethical problems and principles which should be considered when designing research studies. Chapter 4 describes how research problems can be operationalized through study variables, and how null and research hypotheses should be formulated.

Chapter 5 compares and contrasts several research approaches and questions, including experimental and quasi-experimental methods. Chapter 6 contains an extensive overview of the role of internal validity in research studies and threats to internal validity. Chapter 7 systematically describes research designs for randomized experimental and quasi-experimental studies and provides a useful summary table of common aspects of each design. Chapter 8 describes single-subject designs and how they should be operationalized and designed. Chapter 9 relates measurement principles and measurement scales to descriptive statistics and introduces the normal curve and its role in research design. Chapter 10 encompasses sampling and external validity and clearly demonstrates the relationship between sampling and the internal and external validity of a study.

Chapter 11 does an excellent job introducing students to inferential methods, hypothesis testing, and effect size. Students are likely to appreciate the lucid exposition and effective use of graphics to describe these important and fundamental topics. Chapter 12 on design classifications is an excellent chapter that distinguishes between and within group designs in a manner likely to be absorbed by students. A section on diagramming designs is explained clearly step-by-step, and serves to ground the student in this important topic, which is frequently assumed to be understood by authors of other design texts.

Chapter 13 informs students about useful strategies to select appropriate statistical procedures for a given research design. Chapter 14 provides a comprehensive explanation of the *t* test and single-factor ANOVA, which serves to ground the student effectively about these fundamental statistical procedures. A useful schematic representation of when to use post hoc multiple comparisons with a one-way ANOVA is also included, which makes this topic (often a source of derision in students) quite accessible. Chapter 15 includes a useful series of decision trees, which guide the student in identifying research design elements, as well as associated statistical procedure(s).

The limited treatment of linear regression in Chapter 16 needs to be developed further, as it does not adequately inform students about the importance of linear regression and its relationship to correlation and multiple regression. Chapter 17 contains a very well laid out treatment of ANOVA, despite its brevity. Chapter 18 provides a surprisingly clear explanation of the mixed ANOVA approach, the gain score approach, and ANCOVA as a means of statistically analyzing pretest-posttest design studies. Chapter 19 describes multiple regression thoroughly, but the description of the relationship between ANOVA and multiple regression is so brief that this important conceptual relationship is likely to be lost on the student. The rather compact description of MANOVA is clearly written, but is so brief that it is likely to be ''brushed over'' by readers. The same observation holds for the treatment of factor analysis, discriminant analysis, and logistic regression, all of which are given little more than cursory coverage in the chapter.

Chapter 20 contains a lucid and understandable description of the major types of reliability and validity. Chapter 21 documents observational techniques, standardized tests and instruments, personality inventories, attitude scales, questionnaires, interviews, and focus groups. Chapter 22 addresses the issues of confidentiality, peer review, research protocols, Institutional Review Boards, response rates, and scientific misconduct. Chapter 23 utilizes graphics effectively to illustrate the relationships among α, β, and power. Chapter 24 evaluates measurement validity, internal validity, and external validity in turn, with an emphasis on their role with respect to the research validity of a study.

Appendix A provides a comprehensive, detailed Glossary written in layman's language, as well as a sizable listing of what the authors have termed ''confusing terms,'' key distinctions between important concepts. This section should be particularly useful to students who frequently have trouble differentiating these commonly used research terms. Appendix B provides the student with a detailed explanation of the components of a research article. The conversational tone employed by the authors leads the reader step-by-step through the sections of a research article, pointing out major issues critical consumers of research should weigh when reviewing research studies. Appendix C provides clear and step-by-step guidance to students in writing research questions and is a feature that is likely to be welcomed by students.

An important and very useful feature of the text is the incorporation of graphics, tree diagrams, and decision trees that delineate relationships among major topics in a logical, coherent way that makes sense to the reader. These tree diagrams serve as schemata and function to help the student understand the complex theoretical relationships between the research approaches described in the book. Although this approach is not unique in statistical texts of this kind, the authors make unusually effective use of the technique in this volume. Each schema is built upon questions students should ask in evaluating research, which leads to other questions, and possible solution strategies. As stated previously, the effective use of diagrams serves to summarize the most complex content, encapsulating important conceptual and empirical relationships, and providing a cognitive map or schemata that provides the student with a useful conceptual map or overview of the content.

The inclusion of a rating scale to evaluate measurement validity and generalizability of constructs and schematic diagrams that depict the relationship between reliability and validity help students to evaluate the overall research validity of a study. As evidenced throughout the text, the use of these scales and graphics provides a vehicle that enhances students' comprehension and retention of these theoretical relationships.

A welcome and more unusual feature of the text is the authors' acknowledgement of the significance of qualitative methodology, an approach frequently all but overlooked in research design and statistical books. The authors describe the relationship between quantitative and qualitative methodologies, and how they can effectively complement each other, and point out that qualitative methods are not necessarily ''lesser'' than quantitative methods. With this positive emphasis, therefore, it is somewhat disappointing to learn that despite the authors' promise of a significant emphasis on qualitative methods, very little actual space is devoted to qualitative methodology after the topic is introduced in Chapter 2. Indeed, only about two pages are expended on interviewing techniques, and a single paragraph is allocated to focus groups, a qualitative methodology applied frequently in many of the disciplines the text is purportedly geared toward. Clearly, this is not what was promised the reader.

By contrast, a big advantage of the authors' treatment of research design is the ease and comprehensibility with which they describe and distinguish within-subjects from between-groups designs, a distinction many students do not fully comprehend. This confusion often hounds students throughout the entire research design course and confounds their ability to understand complex research designs. The authors take an effective step-by-step approach to describing how data can be arranged, and how these arrangements are related to the research question(s). Using between-groups, within-subjects, and mixed designs to classify research designs, they describe how each research design applies to comparative, experimental, and quasi-experimental approaches. They also show how these three types of designs are related, the types of statistics associated with each design, and the data layouts associated with each design. This systematic and orderly exposition provides students with a conceptual foundation from which to absorb more complex designs and statistical procedures.

Chapter 24, the last chapter of the book, integrates significant points from earlier chapters into a framework for the analysis and evaluation of research articles. Aggregating 16 questions and six rating scales introduced in earlier chapters of the text, the authors provide a systematic guideline students can use to evaluate the research validity or validity

of a whole study. This is an effective tool for students seeking strategies to evaluate research journal articles, and is likely to help students identify key elements to focus their attention.

If you are seeking an excellent overall introduction to research designs and statistical methods that is procedural, clearly written, and which makes effective use of graphics, then this text is likely to fit the bill. Note that it would be judicious to add supplemental materials to address some topics, particularly multivariate statistical procedures and qualitative methods. The authors have succeeded, however, in writing a text that students in many disciplines will find very helpful in learning to interpret and analyze published research results.

Sylvia Kay Fisher
U. S. Bureau of Labor Statistics
Office of Survey Methods Research
Room 4915
2 Massachusetts Avenue N.E.
Washington, DC 20212
U.S.A.
Phone: (202) 691-7382
Fax: (202) 691-7426
e-mail: fisher_s@bls.com

Timothy C. Urdan. *Statistics in Plain English*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2001. ISBN 0-8058-3442-7. 149 pp. 19.95 USD.

The preface of Timothy Urdan's ''Statistics in Plain English'' includes the statement that ''[t]his book is not intended to be used as a primary source of information for those who are unfamiliar with statistics. Rather, it is meant to be a supplement to a more detailed statistics textbook...''. Thus, the topics to be covered, and the amount of detail given to the covered topics, are not well-defined in advance. Presumably the presentation level would not be overly technical, but there still would be latitude as to the level of mathematical background that is assumed. For example, one choice is to assume the typical reader is comfortable with manipulating numbers and thus skip most arithmetic details, while another choice would be to present a substantial amount of detail to allow almost any reader to follow the calculations. The book varies between these choices in the amount of detail that is presented.

The book is not small; the chapters are a total of 133 pages in length, along with a preface, four appendices, an index of terms, and a glossary of symbols (each chapter also concludes with a glossary of terms and symbols). A book of this length might reasonably be expected to contain substantial detail for each of the topics chosen for presentation. This is often true, and explicit disclaimers are made in several later chapters where the author felt that this was not the case.

The topics chosen for discussion are sensible: measures of central tendency and variability, the normal distribution, correlation, *t*-tests, analysis of variance, and regression. There is appropriate emphasis that usually in practice one works with data from a sample,

but the primary interest is to make estimates of, or inferences about, some population from which the sample came.

To draw readers' attention to words and phrases that are listed in the index of terms and defined at the end of the chapter where they first appear, perhaps some distinctive font feature such as bold or italic always should have been used in the book. (This is often, but not always, done e.g., ''boxplot'' on page 14 is in bold font, but ''normal distribution'' on page 3 is not.)

The concept of ''sample'' is introduced in Chapter 1. The statement is made that ''all samples are representative of some population,'' but the definition given for ''sample'' is ''a subset drawn from the larger population.'' Unfortunately, there is not an emphasis that a sample should be a *probability* sample in order to be termed ''representative.'' In Chapter 3, ''sample'' is discussed again, including ''convenience sampling.'' The author does not take the opportunity to describe the obvious limitations of convenience samples, and recommend against their use; instead, the following is stated: ''a sample is selected and its characteristics are noted (e.g., race, socioeconomic status, and unique or special characteristics) so that the population that the sample represents can be inferred.'' This is a serious shortcoming that hopefully will be rectified if future editions of the book are released.

Some other observations on Chapter 1: Table 1.1, which attempts to present a single formula for the mean (sample or population), might have been clearer if two different formulas (sample, population) were presented (as is done, e.g., in Table 2.1 in Chapter 2 for variance and standard deviation). The median is defined for an even number of terms as taking the average of the two central values; a more general definition of ''median'' allows any value between the two central values to be used. There is a discussion on page 3 that could be interpreted to imply that all distributions are either normal or skewed; i.e., no nonnormal symmetric distributions exist. However, it is mentioned in Chapter 5 that the *t*-distributions are symmetric.

Chapter 2 discusses measures of variability. There is a discussion on pp. 8–9 that attempts to justify, in nontechnical terms, why the sample variance uses a divisor of $(n-1)$ rather than n. First, it is mentioned that using any number other than the sample mean in a variance estimate calculation leads to a larger estimate (which is true); the claim is then made that variance estimates using the sample mean instead of the population mean will tend to be too small, relative to the population variance (this also is true, but does not obviously follow from the previous statement). There is a good discussion on page 11 that notes that the wording of questions can influence the responses obtained.

The normal distribution is discussed in Chapter 3. One of the ''characteristics'' presented for the normal distribution is that it is ''asymptotic;'' this term's definition includes the common fallacy of asymptotic graphs ''never touching.''

The theme of Chapter 4 is standardization of data. Appropriate emphasis is given to how standardized data from two samples are more easily compared than unstandardized data. Use of a standard normal distribution table (Appendix A) also is discussed.

Chapter 5 introduces sample-size related concepts such as standard errors, the central limit theorem, *t*-distributions, and degrees of freedom. The introduction of the central limit theorem omits the important word ''approximately;'' i.e., the book states that the mean from samples of size ≥ 30 is normally distributed, rather than stating that the mean from samples of size ≥ 30 is *approximately* normally distributed. (The definition of the central

limit theorem that appears in the glossary at the end of the chapter is more accurate.) Use of Appendix B (t-distribution table) begins in this chapter but never is explained adequately. (There is an example in Chapter 6 that discusses what to do when the degrees of freedom do not match one of the available values in the Appendix B table.) A discussion at the bottom of page 38 would be clearer if it were mentioned that the t-distribution with degrees of freedom is the same as the standard normal distribution (also note that the ''○'' symbol, which appears in this chapter, is not defined until Chapter 6).

The concepts of hypothesis testing, significance, etc. are introduced in Chapter 6. This chapter, and those that follow, include examples that show statistical computer software output (in this book, from SPSS) and helpful discussion of how to decipher what is contained in the output. A clear distinction is made between statistical significance and practical significance. There is mention made of selecting an infinite number of samples of size 1,000 from the (finite) population of U.S. men, and several potential opportunities for fruitful discussions are omitted, e.g.: (1) only a finite number of distinct samples are possible (and how to calculate how many there are); (2) the notion of sampling with replacement, versus sampling without replacement; (3) that there are so many distinct samples that approximating this situation with a continuous distribution is reasonable and appropriate. Confusion could be induced by the discussion on page 45 where the t-distribution with 120 degrees of freedom is said to be identical to the normal distribution, but then the statement is made that probabilities from the two (or one?) distributions are ''virtually'' identical.

Correlation and related measures are introduced in Chapter 7. The presentation of how the correlation coefficient is computed (the author uses z-scores) is good. The important distinction between correlation and causation is stressed.

t-tests are the subject of Chapter 8. It is a disappointment that the concept of degrees of freedom is discussed very briefly, but not fully explained, in this chapter. Given the level of detail devoted to other topics in the book, and the length of the book, devotion of some space to discussing this important concept seems appropriate, to provide some level of understanding as to why various hypothesis test statistics contain divisors/multipliers like $(n-2)$, $(n-k)$, etc.

Three of the final four chapters (9–11) discuss various types of ANOVA (one-way, factorial, repeated-measures). Examples that use Appendices C and D (F-distribution table and studentized range table) occur here. A section near the beginning of Chapter 10 mentions several things that should be checked for (roughly equal number of cases per group, homogeneity of variance) with the statement ''[j]ust as with one-way ANOVA'' (but it was not mentioned in Chapter 9, where one-way ANOVA was discussed, that these things should be checked for). Examples are given in the three chapters that help to point out distinctions between various types of ANOVA. Main effects and interactions are presented and discussed.

The final chapter introduces regression. The strong tie between simple regression and correlation is noted. There is a good presentation of the interpretation of the coefficients in multiple regression.

A new book can be expected to contain some typographical errors, and this book meets that expectation. However, it seems that there are an excessive number of errors in this book; more effort should have been made to proofread the manuscript before it

was printed. A list of all errors found will not be given here. Some could cause confusion, e.g., in Chapter 7, page 63, where the denominator of a test statistic is described to be ''the standard error of the standard error of the sample correlation coefficient'' instead of ''the standard error of the sample correlation coefficient.''

Also, as indicated in several examples given above, there are inconsistencies in the presentation of similar material that appears in more than one place, e.g., results that follow from the central limit theorem. A reader who only reads part of the book might be misled, while a reader who reads the entire book and notes the inconsistencies might be confused.

This book makes a sincere effort to address a well-known deficit in quantitative literacy. If some readers gain helpful insights from it, the book will have succeeded at its goal.

Chris Moriarity
National Center for Health Statistics
Hyattsville, MD
U.S.A.
Phone: 301-458-4384
Fax: 301-458-4031
Takoma Park, MD 20912
U.S.A.
Phone: 301-270-3416
email: chrismor@cpcug.org
or cdm7@cdc.gov

Margaret A.M. Murray. *Women Becoming Mathematicians: Creating a Professional Identity in Post-World War II America*. Cambridge, Massachusetts: The MIT Press, 2000. ISBN 0-262-13369-5. 277 pp. (cloth). 29.95 USD.

Women have always been and still are under-represented in most sciences, engineering, and mathematics. The reasons for this have been debated and explored, particularly with an increase in the importance of these fields in today's technology-driven economy. Margaret A.M. Murray's inspiring book explores the lives of 36 of the 200 women who completed a PhD in mathematics during the 1940's and 1950's, the period with the lowest percentage of women receiving mathematics PhD's this century. She discusses these women's contributions to mathematics, their achievements, struggles, and experiences working towards and becoming mathematicians.

The book compares and contrasts the lives of the 36 women, including their family backgrounds, schooling, graduate work, career development, and their overall identity as mathematicians. Most of the women talked about influences in their life that helped them succeed in a male-dominated field. These influences were encouraging adults in their early life, inspirational teachers and parents, and an overall desire to persevere against the odds. Further, Murray discusses the myth of the mathematical research career. This myth says that a person shows mathematical talent early on and continues to pursue it without a break to a research-driven career. This myth contradicts the experience of most of these particular women's lives, their interests, and their pursuit of a mathematics

PhD. Many did not choose mathematics as their career until later in life and many took breaks from their careers for a variety of reasons.

Murray intermixes the life stories of the women by dividing the book into life stages, and then talking about the 36 women together in each section. This format makes it easy to compare the women's lives, but makes it difficult to link a specific person's experiences with their accomplishments and future. Murray sets the personal biographies against a more general historical background, giving a good context for understanding the events and social expectations that affected these women on their road to becoming mathematicians.

Many of the women experienced varying degrees of sexism throughout their careers, from being told by teachers that women should not pursue mathematics to being expected to serve coffee and pastries to male colleagues. In addition, many talk of receiving lower pay than their male colleagues.

As Murray states in an outside interview about the book, ''none of these women were able to maintain the kind of single-minded focus upon mathematics that characterizes the ideal 'male' career.'' Indeed, when Murray describes the women, she describes their whole person; their many professional achievements, as well as their personal goals and accomplishments.

Women and men interested in history, sociology, and mathematics would find this book entertaining and informative. *Women Becoming Mathematicians* brings about an awareness that there are many ways to be a mathematician. These different ways are highlighted by the women who earned their PhD's at a time when being a woman in mathematics was rare and sometimes looked down upon. These women helped create an environment of increasing tolerance towards females in the mathematical profession.

Kathleen Ott
National Agricultural Statistics Service
United States Department of Agriculture
3251 Old Lee Highway, Suite 305
Fairfax, VA 22039
U.S.A.
Phone: 703-235-5213
Fax: 703-235-5117
e-mail: kott@nass.usda.gov

SDA – Survey Documentation and Analysis.
Vendor: This set of programs is developed and maintained by the Computer-assisted Survey Methods Program (CSM) at the University of California, Berkeley. CSM also develops the CASES software package. (http://csa.berkeley.edu:7502)
Costs: 2,000 USD or it is included free with certain CASES software licensing agreements.

SDA (Survey Documentation and Analysis) is a set of programs that allows a user to create a data archive on the World Wide Web. Once you have successfully created an SDA archive site researchers and students or the general public can come to your web site to browse your data set codebooks or perform various types of statistical analysis. The

current version of SDA allows users to run frequencies, cross-tabulations, comparison of means, and the comparison of correlations.

All the data analysis programs included in the SDA package run directly from within the Web Browser. The software works very quickly, and in most cases the researchers or students can get their results back within seconds, even when using large data sets. The speed at which SDA returns the results to the user's browser window is a result of the way variables are stored on the server. Unlike traditional data bases, where variables are located in a single file, the SDA data structure creates a different file for each variable. This eliminates the need to continually search for variables within a large data base when performing statistical analysis.

Another SDA feature is called ''subsetting.'' This procedure allows you to provide an option for the person using your archive to generate and download a subset of any of your data variables. This procedure will create the data file and produce a codebook for the specified subset, also the procedure gives an option for creating a file containing data definitions for either SAS, SPSS, or STATA.

Installing the software on our web server did involve a few calls to the vendor to get it working. My experience is setting up the software on Microsoft's IIS Web Server. For those people considering setting up the software on a Unix server, the process may be easier since Berkeley's own SDA web site is running on a Unix server. No matter what web server you are using you will need to get cgi-bin access or support from the server administrator to make the software work.

Setting up data files to be archived is a process that is similar to setting up a SAS or SPSS data file. This process basically involves choosing variable names, category labels, missing data codes, variable locations and variable widths. SDA does provide a DOS program that converts SAS, SPSS, or STATA setup files into a SDA ''DDL'' (data description language) setup file. I have found, in converting SPSS setup files, that this program does a good job. However, you will have to review, and likely make some modifications to your DDL file after using the conversion program. Additionally, for most variables you will probably want to add the exact question wording to, your DDL file. This requires inserting the question wording in a text label. While this step is not necessary, having the exact question wording show up in the statistical output and codebook certainly enhances your data archive.

In order to make your data set more user-friendly on the web, there are a few other important steps in addition to creating a DDL file that describes your data. To begin with, you will likely want to create one or two simple ''HTML'' or text file(s) that describe your data set and provide users with information on using data set weights. Next, you will want to create a file that organizes and groups your data variables to make it easier for users to locate variables. After creating these additional files you need to run a program called ''xcodebk'' to create your online codebook and format your additional HTML or text files as frame options in your online codebook. While none of these steps is very difficult to master, it takes some time getting used to the overall procedure. As with working with most software applications, the more you use it, the faster and the easier it is to set up data archives using the SDA software.

To create a data archive on the web you do not need to configure the SDA feature called ''subsetting.'' However, this feature will make your data more useful for those users who

would like to analyze your data using other statistical packages. There are several things to consider in deciding whether to allow subsetting of your data. First, setting up data files for subsetting does not happen automatically, it requires a few extra steps. Second, subsetting can require lots of temporary write space on your server, because download files have to be written on the server. Also, you must remember to regularly remove these temporary files from your web server; it does not happen automatically. Finally, since you are making download data available, you may be setting yourself up for lots of future questions about your data.

One minor frustration using SDA is that the web user needs to cut and paste or remember variable names when choosing to do analysis on the web. With large data sets, the variables and the names become very difficult to remember. The current logic requires that you find the variable name you need in the codebook, making it almost mandatory that you open a second browser window that shows the codebook. A nice feature would be a drop down list of variables on the analysis page. Another area that could be improved is that the procedures for recoding variables take a little time getting used to.

Perhaps the most appealing innovation I found in using SDA on the web is that the results are presented using a clever color shading scheme that makes it easy to interpret your results.

So, who should buy this software? There are three main reasons for purchasing it. First, for teaching statistics and analyzing data in a classroom setting. Second, for providing public access to data sets over the Internet. Third, for people who want to provide data files on their intranet for employees to have easy access. We currently use SDA for all three reasons with the most important use being making data sets publicly available.

SDA makes an excellent teaching tool, since it allows a professor to concentrate on teaching statistics and not have to worry about making sure students get the correct software and manuals. Also SDA is interactive and works from a web browser, and almost all university students have web access. Currently, most statistics courses taught in the social sciences require at least one full class session dedicated to teaching the students how to open their data files, use the software, save their work and print their output. Again, teaching with SDA is easier since most students already know how to print and save html files (all SDA runs and output are html files) and getting access to the data files is done by simply giving the student the correct URL (uniform resource link).

In addition to using data to teach, most universities and government agencies are also committed to providing public access data sets for their campus community or the general public. These data archives are usually in multiple formats and employ several people in answering questions, helping with downloading, and other general maintenance. Providing these data sets in an on-line SDA archive would greatly reduce the amount of duplication and downloading of data files. In addition, it would provide a standard format in which all archived data could be stored.

Finally, an SDA data archive would be useful on a company intranet where data access could still be through the web browser, but limited by the intranet administrator. We put SDA files on an intranet in order to test and work on formatting prior to making them publicly available on the Internet. Also, many large companies, governments or universities may want to restrict the use of on-line data files to their employees or clients. This could be accomplished by setting up an intranet and putting the SDA archive within the intranet.

Overall, SDA is a useful set of programs that provides users with a quick and easy way of generating statistics on large public access data files. The software would greatly enhance data archive sites at universities and government agencies. It is also an ideal solution for any organization that wants to share its survey data with more users.

Tim Triplett
Survey Research Center
University of Maryland
Room 1103
Art Sociology Building
College Park, MD 20742
U.S.A.
Phone: 301-314-7832
Fax: 301-314-9070
e-mail: tim@srcmail.umd.edu
or ttriplett@srcmail.umd.edu
http://www.bsos.umd.edu/src/tim