How to Analyze ANES Survey Data

May 2010

(Revised September 28, 2010)

Matthew DeBell
Stanford University
American National Election Studies

Suggested Citation
DeBell, Matthew. 2010. How to Analyze ANES Survey Data. ANES Technical Report Series no. nes012492. Palo Alto, CA, and Ann Arbor, MI: Stanford University and the University of Michigan.

Acknowledgments
This material is based upon work supported by the National Science Foundation under Grant Nos. SES-0535332, SES-0535334, SES-0937715, and SES-0937727. ANES is also supported by Stanford University and the University of Michigan. This report draws (sometimes verbatim) upon previous documentation of the American National Election Studies.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, Stanford University, or the University of Michigan.

The ANES Principal Investigators are Vincent Hutchings at the University of Michigan and Gary Segura and Simon Jackman at Stanford University. Ted Brader at the University of Michigan is Associate Principal Investigator. The prior ANES Principal Investigators were Jon A. Krosnick at Stanford University and Arthur Lupia at the University of Michigan, with Vincent Hutchings as Associate Principal Investigator.

Thanks to Catherine Wilson for assistance.

Contact
The ANES website address is http://www.electionstudies.org
Any questions not answered on the ANES website or by this report may be directed to ANES staff at anes@electionstudies.org Feedback is always welcome and is invaluable for improving the ANES data and documentation.

CONTENTS

I. THIS GUIDE

This is a "how-to" guide for the analysis of data from the American National Election Studies (ANES). Proper analysis of ANES data is not necessarily hard, but it does require specialized methods that so far have not been widely used by social scientists.

The guide has three preludes and a main feature. The preludes are an overview of the ANES, describing its mission, history, methodologies, and core content; instructions for finding ANES documentation and data; and instructions for preparing an analysis file. The main feature is the section on analysis of ANES data, which describes how to use weights and how to calculate sampling errors and tests of statistical significance that account for the ANES sample design. The guide closes with examples of analysis using the recommended method and shows the contrasts in results with conventional methods.

The Analysis section of this guide is an elaboration of two simple but important statements:

1. **ANES data analysis should be done with weights** for any analysis that is intended to generalize to the population, because unweighted ANES data do not represent the population as well as weighted data.

2. **Significance testing** and associated statistics (standard errors, confidence intervals, *p* values, *t* values, and other statistics related to variance estimation) **should be based on methods that are consistent with the complex sample designs used in the ANES**, because ordinary methods that assume simple random sampling produce biased estimates of significance with ANES data.

This guide is for readers who are familiar with statistics and statistical software, but who may not be familiar with the analysis of complex-sample survey data in general or ANES data in particular.

II. OVERVIEW OF ANES

Mission & History

The mission of the American National Election Studies (ANES) is to advance the scientific study of public opinion and political behavior, specifically by providing high-quality data that measure many variables and support rich hypothesis testing to help scholars understand voter turnout and vote choice.

The ANES traces its origins to a national opinion survey before and after the 1948 presidential election conducted at the University of Michigan's Survey Research Center (SRC) by Angus Campbell and Robert L. Kahn. In every presidential election year from 1948 through 2004 the University of Michigan SRC conducted an election study featuring a pre-election and post-election interview with a representative sample of Americans. In 2006, the ANES became a partnership between Stanford University's Institute for Research in the Social Sciences and the University of Michigan's Institute for Social Research. The latest research projects for the ANES are a continuation of the presidential time series studies and a new 2008-2009 panel study spanning the 2008 election season and the inauguration of the new President.

Methodologies

The main focus of ANES efforts over the decades has been on *Time Series studies* conducted shortly before and after the presidential elections since 1948. ANES has also conducted post-election surveys in most even-numbered non-presidential election years since 1956. Most of these surveys have been conducted using face-to-face interviews in respondents' homes, with between 1200 and 2500 completed

interviews. These studies have repeated many questions over a span of decades, allowing scholars to monitor trends in public opinion over time.

The ANES has also regularly conducted *Pilot Studies* to test and refine new questions for the Time Series. Pilot studies have usually been conducted by telephone. The most recent Pilot Study was conducted by telephone in 2006 with respondents who had previously completed the 2004 Time Series study.

Other major surveys have included panel studies that span individual Time Series studies, stand-alone panel studies (including the 2008-09 Panel Study), and other stand-alone studies such as the Senate Election Studies and the 1982 Methods Comparison Study.

ANES studies use *probability samples*, in which each person in the target population has a known, nonzero probability of being selected for the study. All ANES studies use procedures known collectively as *complex sampling* to distinguish them from simple random sampling. Among the key methods that various ANES studies have employed that make ANES samples complex are the following.

> *Oversampling*: In some ANES studies, members of minority groups, such as blacks and Hispanics in the 2008 Time Series, have been sampled at a higher rate than their proportion in the population. This is done in order to obtain larger numbers of members of these groups in the sample, increasing the precision of statistics estimated for members of these groups.
>
> *Stratified cluster sampling*: Samples for ANES surveys conducted using face-to-face interviews are drawn using geographic areas. A simple random sample of Americans would result in a sample of people scattered all over the country, and visiting each one for a face-to-face interview would be prohibitively expensive. Area sampling involves first drawing a sample of regions of the country, and then sampling successively smaller geographic areas within those regions, before sampling specific households within the sampled small areas. The benefit of this procedure is that it is far less expensive to conduct interviews when members of the sample are clustered together in a limited number of areas.
>
> *Within-household sampling*: ANES studies conducted face-to-face select households within geographic areas. Studies conducted by telephone select households served by a telephone number. In both types of study, one stage of the process involves selecting one individual who lives in a household. If a person lives alone, once his or her household has been contacted, the ANES will seek an interview with that person. In a sampled household with just one resident meeting ANES eligibility criteria (such as being an adult U.S. citizen), there is a 100% chance that ANES will seek an interview with that resident. In a sampled household with more than one eligible resident, ANES will seek an interview with one resident who is randomly selected from those who are eligible. Thus, the probability that ANES will seek an interview with any particular eligible resident is equal to the inverse of the number of eligible residents in the household.

The ANES time series has panel components and cross-sectional components. Each time series study of one presidential election year consists of a pre-election survey and a post-election survey. Respondents to the post-election survey previously completed that year's pre-election survey, so as a study of a single election, each study in a presidential election year is a two-wave panel study. In addition, in some years, ANES re-contacted respondents to surveys from previous years, adding a further panel component to these studies. However, in most studies, from one election year to the next, entirely new samples are drawn. Thus the decades-worth of time series data primarily constitute a set of repeated cross-sectional surveys, not longitudinal surveys.

Core Content

The ANES Time Series studies have featured a core set of questions that have been repeated over time. Exhibit 1 lists topics that have appeared on multiple waves. Some of these are described below.

Partisanship and attitudes towards parties. Party identification is among the most widely used variables in public opinion research. This variable is normally measured using a seven-point scale: Strong Democrat, not very strong Democrat, independent closer to the Democratic party, independent, independent closer to the Republican party, not very strong Republican, strong Republican. The ANES has consistently asked, "Generally speaking, do you think of yourself as a Republican, a Democrat, an independent, or what?" If the respondent answers "Republican" or "Democrat," the followup question asks, "Would you call yourself a strong [Republican/Democrat] or a not very strong [Republican/Democrat]?" If the respondent does not answer "Republican" or "Democrat," the followup question asks, "Do you think of yourself as closer to the Republican Party or to the Democratic Party?"

In addition to party identification, questions about partisanship and attitudes toward parties ask what the respondent likes and dislikes about each party, feeling thermometers for the parties, and whether there are important differences between what the parties stand for.

Candidate and incumbent evaluations. Evaluations of candidates and incumbents regularly include feeling thermometers, perceptions of candidate traits, emotional responses to candidates – whether the candidate makes the respondent feel angry, hopeful, proud, or afraid – and approval ratings of incumbents' job performance.

Issues. Issue coverage includes an open-ended question asking about the country's most important problems and closed-ended questions about social welfare, racial policy, economic issues, foreign relations, social issues, the environment, and federal spending.

Ideology and values. In this category, questions address respondents' religious values, moral traditionalism, and attitudes regarding equality and race, social trust and altruism, cognitive style, and other values and dispositions.

System support questions measure general attitudes toward the political system, the nation, or the government. Questions on trust in government and political efficacy are widely used indicators of system support. Other questions on this topic address the power of the federal government, patriotism, and other aspects of system support.

Political participation and mobilization cover several important topics. Vote choice and voter turnout are focal dependent variables for the ANES. Respondents' turnout and vote choice in the prior election, the current primary, and the current general election are asked. During the pre-election questionnaire, the respondent's vote intention is asked. Other dimensions of participation include interest in politics, discussion of politics, political knowledge, contacts and direct involvement with the political campaigns, and other civic participation.

Media. Respondents are asked about their use of news media, such as television, the Internet, newspapers, and radio.

Social groups. Feeling thermometer questions are asked for social groups, and respondents report their own sense of identification or closeness to social groups.

Personal and demographic data cover a broad range of information: educational attainment employment status, religious identification, race/ethnicity, date of birth, mobility of residence, income, social class, and other variables including labor union membership, home ownership, and marital status.

Exhibit 1. Recurring question topic list for the ANES Time Series

I. Partisanship and attitudes towards parties
- Feelings about parties
- Party identification
- Closeness to parties
- Party performance
- Role of parties

II. Candidate and incumbent evaluations
- Feeling thermometers
- Candidate evaluations
- Traits of president, candidates
- Affect toward President, candidates
- Evaluation of Congressional candidates

IIA. Performance evaluations
- Performance of current president
- Retrospective evaluations
- Other government performance
- Candidate performance
- Congressional candidate performance

IIB. Contact with Congressional candidates/incumbents

III. Issues
- Social welfare issues
- Racial policy issues
- Economic issues
- Foreign relations
- Social issues
- Environment
- Federal spending

IV. Ideology and values
- Religious values
- Moral Traditionalism
- Equalitarianism and race
- Social trust and altruism
- Cognitive style
- Other values and predispositions

V. System support
- Trust in government
- Power of the federal government
- Efficacy and gov't responsiveness
- Patriotism
- Other system support

VI. Political participation and mobilization
- A. Civic participation
- B. Engagement and information
  - Interest
  - Discuss politics
  - Political knowledge
- C. Campaign activity
  - Respondent's campaign activity
  - Campaign contact/solicitation
- D. Registration, turnout, vote choice
  - Previous Presidential election
  - Vote intention
  - Turnout & registration
  - Vote choice
  - State primary/caucus

VII. Media
- Use of media

VIII. Social groups
- Group thermometers
- Group identification and closeness

IX. Personal and demographic data
- Education
- Employment status
- Religious identification
- Race/ethnicity
- DOB
- Mobility
- Income
- Social class
- Other personal and demographic data (labor union, home tenure, marital status, etc.)

III. FINDING ANES DOCUMENTATION & DATA

Documentation

"Documentation" comes before "Data" in the title of this section because you should read the documentation first. Documentation and data for each ANES study may be downloaded from the ANES web site.

To see a description of a study as well as links to the study's errata, questionnaires, codebook, and other study-specific resources, go to http://www.electionstudies.org. Click on the Data Center to see the list of studies and click on the Study Page for the study that interests you.

Errata
All survey data have errors. Before analyzing any data file, look for documentation of the errors. ANES documents errors on the ANES web site, and the information is often important to know before you begin an analysis. From the ANES Data Center on the web page, click the Errata link associated with any dataset to see the list of known errors in that dataset. For example, in the 2004 dataset, as of the time of this writing, an erratum note says to drop case 357 from the data file because the person who completed the interview was not properly sampled. (At some point in the future, a new data file that omits case 357 may be posted to the web site.)

Questionnaires
ANES questionnaires show the exact wording of each question asked to respondents, as well as the categories and codes used to record their answers and, in some cases, the names of the variables where the answers are recorded.

Questionnaires also document the order in which questions were asked and show which respondents were asked each particular survey question. Questions were asked in the same order in which they appear on the questionnaire unless the questionnaire contains a note indicating that the order was changed, such as noting that the order of a set of questions was randomized. Each question was posed to all survey respondents unless a note specifies conditions in which to ask a particular question. For example, a note that says "(IF YES)" means that a question was asked only if the respondent answered "yes" to the previous question.

Because ANES has conducted a large number of different surveys over several decades, with different studies managed by different generations of Principal Investigators and professional staff, the format of the questionnaires is not consistent. Some questionnaires include question-by-question documentation (called QxQs) that was provided to interviewers during their training and was available to interviewers during the interviews for their reference, to help them administer the questionnaire correctly.

Questionnaires that date from the ANES era before interviewers used laptop computers are typically scanned images of a paper questionnaire form. Newer questionnaires may be PDF or html files that document the questions that were displayed on a computer.

Codebooks
Codebooks for most ANES studies are divided into three parts:

- *Introduction*: the codebook introduction describes the study's design, sample, response rate, procedures, and content. The introduction also includes a list of all the variables on the public-use data file in the order in which the variables appear on the file. The list of variables includes a summary of each variable's content. These summaries are written tersely and do not necessarily reflect the exact wording of the question. For the question wording, see the questionnaire or the "Variables" section of the codebook.
- *Variables*: the variables section of the codebook documents each variable by listing the question topic, the wording of the question asked, interviewer instructions, codes used for responses or missing data, and notes about the variable, if any.

- *Appendices*: the codebook appendices include information that does not fit in the introduction or variable list. Usually the appendices contain codes used for variables that have a very large number of codes that would be unwieldy to include in the variable list, such as codes for industry and occupation. Several ANES codebook appendices also contain instructions for design-consistent variance estimation. (For more on this, see the Analysis section of this guide.)

Other Study-Specific Documentation
Additional resources are often available for ANES studies. These include interviewer instructions, planning committee reports, and respondent booklets and so-called "show cards," which are materials handed to the respondent during the face-to-face interview to help the respondent understand the question or the response options. For recent studies, additional documentation may include extensive and detailed reports on study methodology.

Pilot Study Reports & Technical Reports
Pilot studies are intended to test new questions for possible inclusion on later ANES studies. Pilot Study reports describe the results of these tests. More than 100 Pilot Study reports written since 1978 can be found at http://www.electionstudies.org/resources/papers/pilotrpt.htm, or by navigating from the ANES home page to the link for Reference Library, and from there to the link for Pilot Study Reports.

The Reference Library page also contains a link to the ANES Technical Reports. Dozens of technical reports address methodological topics such as the response rate, survey mode, bias, weighting, and vote validation.

Datasets

From the home page at http://www.electionstudies.org, click on the Data Center to see the list of studies and click on the "Download Dataset" link to download a zipped folder containing the data and its codebook. Be sure to also follow the link to the Study Page to review the study's questionnaire and errata.

Creating an Analysis File

How to Get the Data
ANES data are freely available for download from the Data Center at www.electionstudies.org. From the home page, click on Data Center to see the list of public data products. These are organized by study year, from 1948 through the most recent study.

Restricted-Use Data
Most ANES data are on these public files, but access to data that may pose a risk to respondent privacy or confidentiality is restricted. Chiefly, restricted data consist of detailed geographical data such as the respondent's ZIP code, occupation data that describe the respondents' occupation and industry using very detailed classification categories, and text that includes interviewers' transcriptions of respondents' answers to open-ended questions. Access to the restricted data may be requested through the Restricted Data Application Process, described at http://www.electionstudies.org/rda/anes_rda.htm . ANES has estimated that the average marginal cost of the activities required to process a restricted data request is $335, excluding overhead. Therefore there is an administrative fee of $335 for a Restricted Data request. As a rule, all such requests from researchers who can provide adequate assurance that they will safeguard the data and are performing bona fide research will be honored.

What You Get When You Download Data
Each dataset is packaged in a ZIP file that contains the data file, documentation, and data definition programs to read the data with statistical software. These files include some or all of the following:

- *Flat data file*. The data are recorded in a plain text file that can be read by statistical software using data definition programs (sometimes called "extract files"). ANES provides these definition programs for SPSS, SAS, and Stata. The flat file is in a plain-text format, described in the section below, Running the Data Definition Programs.

- *SPSS data definition program*. When you run the SPSS data definition program in SPSS, the SPSS software will read the flat data file and create a data file in SPSS format. See Running the Data Definition Programs, below, for instructions.
- *SAS data definition program*. When you run the SAS data definition program in SAS, the SAS software will read the flat data file and create a data file in SAS format. See Running the Data Definition Programs, below, for instructions.
- *SPSS Portable file*. All datasets are provided as SPSS Portable files. These are data files that maximize compatibility with different versions of SPSS and can be shared by versions of SPSS that run under different operating systems.
- *SAS Transport file*. Some ANES ZIP packages contain transport files for SAS. Like SPSS portable files, these are data files that maximize compatibility with different versions of SAS and can be shared by versions of SAS that run under different operating systems.
- *Stata data definition program*. Stata will read the flat data file and create a data file in Stata format when you run this program.
- *Codebook introduction*. A codebook introduction describes the study and its methodology.
- *Codebook variable documentation*. The codebook variable documentation lists each variable on the file and includes question wording and valid values for the variables.
- *Codebook appendix*. Appendices provide additional details about the dataset such as information about the coding of open-ended variables.
- *Readme.txt* file. This file contains a description of the files included in the ZIP file and may contain other important notes about the data release.

Running the Data Definition Programs

ANES data files are provided as "flat" files, in a plain-text format. Statistical software such as Stata, SPSS, and SAS can convert these flat files into their own data formats if you give the software instructions on how to read the file.

These text files are formatted with a single row of data for each respondent. The first row of data contains the variable names, and subsequent rows represent individual survey respondents, with a separate row for each respondent. Columns represent variables. Individual variables are "comma-delimited," meaning they are separated by commas. This format is also known as "comma-separated values," or CSV, and is a common database format. Variables are also "fixed widths," meaning that an individual variable occupies the same number of characters for every respondent.

The following is an example of a fragment of the 2006 ANES Pilot Study flat file, showing the first few characters from the first four lines of the file.

```
Version,V06P001,V06P002,V06P003,V06P004,V06P005,V06P006,V06P007a,
ANES_2006pilot_VERSION:2007-Apr-26,   1, 0.3376,1,1,1,52, 3,
ANES_2006pilot_VERSION:2007-Apr-26,   2, 0.4221,1,1,1,49, 2,
ANES_2006pilot_VERSION:2007-Apr-26,   3, 1.1993,1,1,2,39,21,
```

The first line lists the names of the variables, separated by commas. The second line is the data for the first respondent listed on the data file, with the data for each respective variable separated by a comma. The third line is the second respondent. The file continues for another 1209 lines, but only the first four are shown here for illustrative purposes.

On most studies, each variable name starts with the letter V, followed by two numbers for the study year, followed by four numbers to uniquely identify the variable in the sequence in which the variables are listed on the file. Thus the variable V080134 would be the 134th variable on the 2008 file. In the 2006 Pilot Study, the 4th character in the variable name is the letter P, to identify pilot data.

The text reading "`ANES_2006pilot_VERSION:2007-Apr-26`" is the value of the first variable (Version) for the first respondent. The comma after this string of text indicates the end of the record for the first variable. The second variable, V06P001 occupies the next four characters and has the value of 1 for

the first case. (V06P001 is the Case ID, and cases are numbered consecutively from 1 through 1,211.) The third variable is V06P002, the weight, and occupies the next seven characters (with values of 0.3376, 0.4221, and 1.1993 for the first, second, and third cases, respectively). This excerpt illustrates only the initial characters of each line. In the actual flat data file, each line continues to the right for a few thousand characters, through variable V06P822b.

To prepare the data files for analysis, follow these steps after downloading the data:

1. When you download the data, the downloaded ZIP folder will contain a file called readme.txt that lists the files included in the download and will name the data file. Read the readme file. Note the name of file containing the plain text data. Data file names usually end in ".dat" or "…dat.txt".

2. Save the plain text data file to your computer and note the path where you save it.

3. The readme file will also specify the names of the data definition programs. Find the name of the data definition program for your preferred statistical software (Stata, SPSS, or SAS), and open that program in the statistical application.

4. In the data definition program, read the instructions and notes at the beginning of the file, if any.

5. In the data definition program, find the part(s) of the program that list the path(s) of the plain text data file and edit the path to specify the file location and file name where you saved the plain text data file in step 2, above. The program may also specify a path and file name where the statistical software will save the new Stata, SPSS, or SAS data file that you will create. Edit this path so that the software saves the new ANES file with a name and location that you want.

6. Run the data definition program in the appropriate statistical software. When the program has finished (which can take a few minutes, depending on the size of the data file and the speed of your computer), check the output to make sure no errors were reported.

Another way to create an analysis file is to have your statistical software read the text file automatically. If you are using SPSS, SAS, or Stata, it is usually better to use the ANES-provided data definition statements, because these statements will provide labels for each of your variables. However, some software (including SPSS) can import data automatically, after you provide basic information about the file described earlier in this section, such as indicating that the file is comma-delimited, has variable names in the first row, and has one case per row beginning on the second row.

Checking the Data File

After creating an analysis file, it is a good practice to quickly check the data file to be sure that it downloaded correctly and was read into your statistical software correctly.

Data integrity checklist:

- Verify that the number of cases on the data file matches the number of cases listed in the codebook. For example, the 2006 Pilot Study survey was completed by 675 respondents, so your analysis file for that dataset should contain exactly 675 cases.
- Verify that the variables on the file are the same as the variable names listed in the codebook. ANES data files often contain hundreds of variables, making it impractical to verify the presence of every variable on the file, but it is wise to verify that the first and last few variables on the data file match the first and last few variables in the codebook.
- For a few selected variables, verify that the frequencies obtained through your statistical software match those listed in the file's documentation. If the documentation provides weighted and unweighted frequencies, match both estimates.

If these checks of data integrity turn up anything amiss, something probably went wrong with the data definition program. Delete the data file and data definition program, download the data again, and run

the data definition program again. If you cannot determine the cause of the problem after looking carefully at your files, contact the ANES staff for assistance at anes@electionstudies.org.

Merging Data Files

There are two ways it can be useful to merge data files. Merging can yield a data file with more respondents (cases) for a given set of variables, or with more variables for a given set of respondents. Or, it can yield both of these at once.

Adding respondents (cases). One of the defining features of the ANES time series datasets is the repetition of the same questions on different surveys over many years. Although different respondents answer the time series surveys in different years (so the time series is not what is normally called a longitudinal study), the same questions are asked of fresh samples each year. If your research would benefit from being able to analyze these data all at once, you can merge two or more data files. Note that in two different data files, a question will typically have two different variable names, even if the question wording was the same.

Adding variables. Some ANES datasets contain data from the same respondents as other ANES datasets, affording an opportunity to merge the files. The 2006 Pilot Study is such a dataset, as the 675 respondents to this survey all completed the pre-election wave of the ANES 2004 Time Series study (not to be confused with the ANES 2004 Panel Study). Adding variables means that for a given set of respondents, you are adding answers to more questions, or adding other information about those respondents, such as new weights or contextual data.

Data merging example: ANES 2004 Time Series + ANES 2006 Pilot Study

There are two distinct ways that these datasets can be merged: 1) starting with the 2004 dataset, the variables from 2006 can be added to the appropriate cases; 2) starting with the 2006 dataset, the variables from 2004 can be added to all of the cases. The first procedure will result in a data file containing 1,212 cases (that is, responses from 1,212 people who completed a pre-election survey in 2004). Among these 1,212 cases, 675 contain additional data from the 2006 survey. The second procedure will result in a data file containing 675 cases, all of whom completed the 2006 survey as well as the pre-election survey in 2004. The former approach is the most versatile and is appropriate if you want to be able to run analyses on the 2004 data or if you want to analyze nonresponse in 2006 by comparing the 2006 respondents to the 2006 nonrespondents. The second approach may be desirable for analyses focused exclusively on the 2006 respondents, because it results in a data file containing only the 2006 cases. The instructions below show both methods, using SPSS. Analogous procedures are available in other software such as Stata and SAS.

As with most SPSS procedures, data file mergers may be accomplished using syntax or menus. Any scientific analysis of data must be replicable, and the best way to assure that a statistical analysis can be repeated – and the best way to assure that you can figure out exactly what you have done if you need to reexamine the work in the future – is to use syntax for all of your work. SPSS has a useful "Paste Syntax" command that allows you to use menus to generate syntax if you do not know how to type a command.

With any data merger, the first two steps are to prepare the two data files. If necessary, download the files and import them to your statistical software, and check them to assure that the cases and variables were imported properly. (See the sections above, How to Get the Data, Running the Data Definition Programs, and Checking the Data File.) If variables are to be added, assure that there is a common ID variable on both files, so that the new data can be matched, and assure that both files have been sorted by that ID variable, so the records are in the same order.

The example SPSS code below merges the 2006 Pilot Study and the 2004 ANES time series datasets twice, first creating a 2006 dataset that adds the 2004 variables to all 2006 cases, and then creating a 2004 dataset that adds the 2006 variables to the available cases. The lines preceded by an asterisk are documentation; SPSS ignores these lines. For this code to work, you need to have saved the 2006 Pilot Study SPSS data file as c:\anes\anes_2006pilot.sav and you need to have saved the 2004 time series

SPSS data as c:\anes\nes2004.sav. If you've saved your SPSS files to other locations, just change the path wherever it appears in the code below. Analogous procedures apply to other ANES datasets.

```
* Purpose: to merge the 2004 NES and 2006 ANES Pilot Study datasets by, first,
creating a file that adds the 2004 variables to the 2006 file, and, second, creating
another file that adds the 2006 variables to the 2004 file.

* create a common ID variable in both files, sort the files by that ID variable,
rename the file version to preserve it after the merge, flag the pilot cases, and
drop case ID 357 from the 2004 data.

get file = 'C:\anes\anes_2006pilot.sav'.
string version06 (A25) .
compute version06 = version.
compute ID = v06p001.
compute pilot=1.
sort cases by ID(A).
save outflie = 'C:\anes\temp_merge06.sav'.

get file = 'C:\anes\nes2004.sav'.
string version04 (A24).
compute version04 = version.
compute ID = v040001.
* the erratum to the 2004 ANES dated Dec 13, 2007, says that case 357 should be
dropped from the data file, so the next line of code drops that case.
select if (v040001 ne 357) .
sort cases by ID(A).
save outfile = 'C:\anes\temp_merge04.sav'.

*merge the files RETAINING ONLY THE 675 PILOT CASES .
MATCH FILES /TABLE=*
 /RENAME (Version = d0)
 /FILE='C:\anes\temp_merge06.sav'
 /RENAME (Version = d1)
 /BY ID
 /DROP= d0 d1.
EXECUTE.
save outflie = 'C:\anes\anes06final_with04vars.sav'.

* merge the files ADDING THE 2006 PILOT DATA TO THE 2004 DATA AND RETAINING ALL 1211
CASES.
* A new variable, status, is coded 1 if the case completed the 2006 Pilot Study and
0 if the case did not complete the Pilot Study.
get file = 'C:\anes\temp_merge04.sav'.
MATCH FILES /FILE=*
 /RENAME (Version = d0)
 /TABLE='C:\anes\temp_merge06.sav'
 /RENAME (Version = d1)
 /BY ID
 /DROP= d0 d1.
EXECUTE.
compute status=0.
if pilot=1 status=1.
save outflie = 'C:\anes\anes04-06final.sav'.
```

IV. ANALYSIS

Summary

ANES surveys use "complex" sampling techniques. To represent the population properly, analysts need to do two things in nearly all analyses:

- Use the weights.
- Calculate sampling errors and significance tests using the Taylor Series method or another design-consistent approach.

This section details these design-consistent methods for analysis of ANES data. The last section of the document provides examples of such analysis.

Conceptual Rationale for Design-Consistent Estimation

ANES data require special analysis techniques because the respondents are not selected using a simple random sample. In a simple random sample, each member of the population has an equal probability of selection that is not affected by the selection of any other member. In the ANES, each member of the population has a theoretically knowable probability of selection, but some population members are more likely to be selected than others. In ANES surveys conducted by random-digit-dialing on the telephone, people who live in households with two telephone lines are twice as likely to be selected as those who live in households with one telephone line. In ANES surveys conducted face to face, the cluster design means that people living near one another are sampled in groups, and these people who live near one another are likely to resemble one another in other ways. In all households, the ANES practice of selecting one household member means that people who live in large households are less likely to be selected than people who live alone or in small households.

Two statistical techniques adjust for these complex aspects of the ANES design. The first is weighting the data and the second is an adjustment to the way we calculate standard errors and statistical significance to produce design-consistent estimates. Weighting accounts for unequal probability of selection, and also helps correct random and systematic errors due to nonresponse. The special techniques for calculating standard errors account for the clustering of the sample. Together, these techniques are called "design-consistent" estimation.

Software

Proper analysis of ANES data requires the use of software that can weight the data and produce design-consistent estimates. Six examples of software that can do this are as follows:

- *Stata*. For information about Stata, see http://www.stata.com
- *SPSS with SPSS Complex Samples*. The SPSS Complex Samples module, in addition to SPSS Base, is required to conduct design-consistent estimation using survey data with complex samples. For information about SPSS, see http://www.spss.com
- *SAS*. For information about SAS, see http://www.sas.com
- *SUDAAN*. SUDAAN can be used as a stand-alone program or can be run from within SAS. For information about SUDAAN, see http://www.rti.org/SUDAAN
- *AM*. AM is free. It does not have very powerful tools for recoding and it does not support the use of syntax, though commands may be saved and re-run later. For information about AM, see http://am.air.org
- *R*. Unlike the five programs described above, R is not a conventional software application. Rather, it is a free, open-source statistical programming language. Programmer-analysts interested in R may wish to consult Crawley (2005), Crawley (2007), and http://www.r-project.org

Users may also be interested in the Survey Documentation and Analysis programs (http://sda.berkeley.edu/) developed and maintained by the Computer-assisted Survey Methods Program (CSM) at the University of California, Berkeley. CSM provides online analysis software that can be used to analyze ANES data and other data for free. ANES does not develop or maintain this resource.

Weights

For the datasets on which they are provided, weights correct for unequal probability of selection and for nonresponse to the survey. This makes the survey's representation of the population more accurate than it would be without the weights. Therefore, always use the weights if you wish to generalize results to the population.

Example: To analyze 2006 Pilot Study data using the weights in SPSS, use the following command:

```
weight by v06p002.
```

After this command is run, all subsequent procedures will be run with weights. To turn the weights off – not that you should – use the following command:

```
weight off.
```

Note: ANES datasets prepared prior to 1988 often do not include weights. In the early years of the ANES, statistical software to analyze weighted data was not widely available. After such software became more widely available, ANES staff often concluded that although the use of weights was technically appropriate, the benefits of weighting were too small to justify the added complexities of weighting. However, for any ANES data file for which weights are included, they should be used. For more information about how ANES weights are created, see the codebook for the relevant survey. For information about the recent calculation of weights in ANES studies see DeBell & Krosnick (2009).

Frequently Asked Question: Do I still need to weight if I'm doing regression?
Most likely yes. Analysts reasonably question whether weights are necessary in multiple regression analyses, provided that variables such as age and educational attainment, which are poststratification components of the weights, are included as independent variables in the regression equation. ANES weights also account for features of the sample such as the probability of selection due to geographic area sampling, within-household sampling, or the number of telephone lines in the household, that are not represented in variables that analysts can or will include in regression analyses. Therefore, to generalize about the population, *always use the weights*, unless you are developing a sophisticated model-based alternative to weighting. Model-based approaches can be a legitimate alternative to using weights, but the difficulty of correctly specifying a design model for a complex-sample survey like ANES can be substantial.

How to Choose the Correct Weight Variable

ANES Time Series datasets include a Pre-election weight and a Post-election weight. Use the weight from the latest wave of the study that furnishes any variables you are analyzing. If all of the variables in your analysis are from the Pre-election survey, use the Pre-election weight. If any or all of your variables are from the Post-election survey, use the Post-election weight.

The reason for this weight selection approach is that the Pre-election weight is optimized for the full sample – for an analysis of all respondents – while the Post-election weight is optimized for the subset of the sample that completed the post-election interview. Thus, you should use the Pre-election weight when your analysis includes all respondents (i.e., it is limited to Pre-election variables) and you should use the Post-election weight when your analysis is limited to the Post-election respondents (i.e., when it includes any variables from the Post-election survey).

For other studies with multiple weights, such as panel studies, see the study's documentation for guidance on weight selection.

Standard Errors & Significance Testing

Optimal Procedure
To obtain correct standard errors and significance tests, set up your software to use Taylor series estimation. The command to do this will specify three variables: the weight, the stratum, and the primary sampling unit or PSU (also sometimes referred to in various sources as a cluster or as a standard error calculation unit or SECU). In the 2006 Pilot study, the codebook shows that the weight is v06p002, the stratum variable is v06p007a, and the SECU is v06p007b.

In Stata, the command to set up any dataset that has been opened in the program is as follows, where WGT is the name of the weight variable, STRAT is the name of the stratum variable, and CLUSTER is the name of the PSU variable:

```
svyset [pweight=WGT], strata(STRAT) psu(CLUSTER)
```

You will need to replace WGT, STRAT, and CLUSTER with the names of the weight, stratum, and cluster variables in the ANES data file you are analyzing. For example, the Stata command to set up the 2006 ANES Pilot Study for design-consistent estimation is as follows:

```
svyset [pweight=v06p002], strata(v06p007a) psu(v06p007b)
```

Similar commands specify the weight, stratum, and cluster variables in other statistical software. In SPSS, the Complex Samples package requires creating an "analysis plan" file that specifies this kind of information. Use the SPSS menus to specify the parameters and SPSS will create a "plan" file like this one:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<SPSSComplexSamples version="1.0">
	<Header copyright="Copyright (c) SPSS Inc., 2006. All Rights Reserved."/>
	<AnalysisDesign SRSestimator="wr" numberOfStages="1">
		<AnalysisStage estimationMethod="wr" stageNumber="1">
			<StrataVarList numberOfVariables="1">
				<Variable name="V06P007a"/>
			</StrataVarList>
			<ClusterVarList numberOfVariables="1">
				<Variable name="V06P007b"/>
			</ClusterVarList>
		</AnalysisStage>
		<Weight>
			<Variable name="V06P002"/>
		</Weight>
	</AnalysisDesign>
</SPSSComplexSamples>
```

Note: in a sample design such as ANES, it normally does not matter whether you treat the design as if it was conducted "with replacement" or "without replacement."

See the documentation for your preferred software for more information.

Creating Strata and PSU Variables in ANES Datasets
The 2006 Pilot Study is the first ANES dataset for which the stratum and PSU are provided as separate variables. In ANES time series studies conducted from 1990 through 2004, the strata and PSUs were

provided in one variable that must be split into two prior to running a svyset command in Stata or an equivalent command in most other software.

The first two digits are the stratum and the last digit is the SECU.

In the 2004 ANES, v040103 contains this information. The stratum is the first two digits of this variable. The SECU is the last digit. In SPSS, this can be bifurcated into stratum and secu variables using a brute force approach with code that begins as follows:

```
if v040103='011' stratum=1.
if v040103='011' secu=1.
if v040103='012' stratum=1.
if v040103='012' secu=2.
if v040103='021' stratum=2.
if v040103='021' secu=1.
* Et cetera, for another few dozen lines of code.
```

More tersely, the same result is obtained when SPSS is commanded as follows:

```
compute stratum = trunc(v040103/10).
compute secu = v040103 - (10*stratum).
```

The same task can be accomplished with analogous code in other software such as SAS or Stata.

Analysis Commands with the Optimal (Taylor Series) Method
After you have defined the weight, strata, and PSUs for your software, use special commands to produce design-consistent estimates using the Taylor Series method. This section provides examples of these commands for selected software. For more details, see the software's documentation.

Stata Commands
In Stata, the "svy" commands will produce design-consistent estimates, provided the svyset command has previously been run. Svy commands merely precede normal Stata commands with the text "svy:".

For example, to run a logistic regression with the dependent variable 'vote' and the independent variable 'party', the standard command (assuming simple random sampling) in Stata would be as follows:

```
logit vote party
```

To produce design-consistent estimates using Taylor series (after having run the svyset command) in Stata, use the svy command, as follows:

```
svy: logit vote party
```

Other Stata commands will produce percentages, crosstabs, means, and other statistics. Examples for percentages (proportions) and crosstabs are shown below:

```
svy: proportion vote party
svy: tab vote party, se deff
```

For additional information about Stata commands to produce design-consistent estimates, consult the software documentation.

SPSS Commands
In SPSS Complex Samples, a CSPLAN file (described above) must be created before analysis to identify the weight, strata, and PSU variables. Once this file has been created, SPSS commands refer to that file. The three command sets below respectively produce frequencies, crosstabs and a logistic

regression for the variables vote (dependent in the regression) and party (independent in the regression). Generally it is easier to use the SPSS menus to generate the syntax than to type it, since this syntax is verbose (especially compared to Stata's).

```
cstabulate
/plan file = "C:\ANES\CSPLAN_ANES06"
/tables variables = VOTE PARTY
/cells tablepct
/statistics se deff
/missing scope=table classmissing=exclude.

cstabulate
/plan file = "C:\ANES\CSPLAN_ANES06"
/tables variables = VOTE by PARTY
/cells rowpct colpct tablepct
/statistics se deff
/missing scope=table classmissing=exclude.

CSLOGISTIC  VOTE(HIGH) BY PARTY
  /PLAN FILE='C:\ANES\CSPLAN_ANES06'
  /MODEL PARTY
  /INTERCEPT INCLUDE=YES SHOW=YES
  /STATISTICS PARAMETER EXP SE DEFF
  /TEST TYPE=F PADJUST=LSD
  /MISSING CLASSMISSING=EXCLUDE
  /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-
006 RELATIVE] LCONVERGE=[0] CHKSEP=20 CILEVEL=95
  /PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

SAS Commands
In SAS, the definition of the weight, strata, and PSU is included in each analysis procedure. The code below produces frequencies and crosstabs and runs a logistic regression on the variables vote and party. (SAS data definition statements are omitted from this example.) The variable names are capitalized only to distinguish them from the rest of the SAS code.

```
proc surveyfreq;
stratum STRATUM;
cluster PSU;
weight WEIGHT;
table VOTE PARTY VOTE*PARTY ;
run;

proc surveylogistic;
stratum STRATUM;
cluster PSU;
weight WEIGHT;
model VOTE = PARTY ;
run;
```

Other Software
For information about commands to produce design-consistent estimates in other software, such as AM, SUDAAN, and R, consult your software documentation.

How to Subset the Data when Using the Taylor Series Method

Researchers often want to analyze only a subset of the data (or a subpopulation), such as by including voters and excluding nonvoters from the analysis. The Taylor series procedures rely upon all of the cases on the data file to arrive at correct estimates, including cases that may be excluded from a particular analysis. Therefore, if you want to run an analysis using anything less than the full sample, you should not simply drop unneeded cases from the data file (temporarily or permanently) prior to an analysis. Instead, you should use your statistical software's subset commands that are intended for use with Taylor Series procedures.

To do this, you can begin by creating a dummy variable, "flag," coded 0 if a case is to be excluded and coded 1 if the case is to be included. This selection variable cannot be missing for any respondent.

In an ordinary, simple-random-sample data file, you could select flagged cases for analysis with a command such as the following in SPSS, which would delete all cases from the data file for which flag is not equal to 1:

```
select if flag = 1 .
```

If you save the file after running this command, cases for which flag is not equal to one are permanently deleted. Do not do this with ANES data if you wish to produce design-consistent estimates. Instead, subset the data file.

Subsets in Stata
Consider the following example Stata code:

```
svy: proportion age, subpop(flag)
```

The "svy: proportion" command returns the percentage of responses for each value of a variable. It requires that the data file have been already set up for design-consistent estimation using the svyset command. (See the Standard Errors and Significance Testing section above for more on svyset.) This example generates proportions for the variable "age", and does so only for the subpopulation of cases identified by the variable "flag". The variable that identifies the subpopulation must be coded 0 for cases that are excluded and 1 for cases that are included.

Other Stata commands use the same kind of syntax to indicate subpopulations.

Subsets in SPSS
Subpopulation options exist in the complex sample menu commands for SPSS complex sample procedures. Specify the subpopulation indicator variable and the value of that variable that indicates cases to be included in the analysis. This is an example of the code to produce design-consistent frequencies for the subpopulation designated by the variable FLAG.

```
cstabulate
/plan file = "C:\ANES\CSPLAN_ANES06"
/tables variables = VOTE PARTY
/subpop table = FLAG display=layered
/cells tablepct
/statistics se deff
/missing scope=table classmissing=exclude.
```

The output from this code will repeat the cstabulate command for each category designated by FLAG.

Subpopulations for regression work slightly differently. You specify the flag variable and also specify which value of the flag indicates cases will be included in the analysis. This is called the "domain variable." Example code is below to run logistic regression that includes cases for which FLAG=1.

```
CSLOGISTIC  VOTE(HIGH) BY PARTY
  /PLAN FILE='C:\ANES\CSPLAN_ANES06'
  /DOMAIN VARIABLE=FLAG(1)
  /MODEL PARTY
  /INTERCEPT INCLUDE=YES SHOW=YES
  /STATISTICS PARAMETER EXP SE DEFF
  /TEST TYPE=F PADJUST=LSD
  /MISSING CLASSMISSING=EXCLUDE
  /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-
006 RELATIVE] LCONVERGE=[0] CHKSEP=20 CILEVEL=95
  /PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

Subsets in other software
Analogous commands are used in other software. See the software manuals for instructions.

Standard Errors on subsets of the dataset that have empty strata. (Or, "how to troubleshoot when subsetting produces empty strata.")
Sometimes attempts to calculate sampling errors on subsets of the dataset do not work because the subset has resulted in a working set of cases that has an empty stratum, empty cluster, or too few cases in a stratum or cluster to complete a complex analysis. In this event, normal Taylor series procedures to calculate standard errors do not work, and most software will return no statistics for sampling error.

In this event, see your software documentation to see if it incorporates an automated procedure to deal with this situation, and to see if you want to use that procedure.

If your software does not provide procedures to deal with this situation, or does not provide procedures that you want to use, follow these step-by-step instructions to collapse strata:

1. Subset the data for the cases you are looking at for your analysis and, for this subset, crosstabulate the stratum variable (v06p007a in the 2006 Pilot Study) by the cluster variable (v06p007b in the 2006 Pilot). If the subset variable is flag1 and cases are selected for analysis if flag1=1, then the code in SPSS would be as follows:

```
temporary.
select if flag1=1.
cross v06p007a by v06p007b.
```

2. Look at the output to identify the cells in this matrix with too few cases. Zero is always too few. One, two, or larger numbers may also be too few, depending on the analysis.

3. Recode stratum (v06p007a) into a new variable and combine adjacent strata to eliminate cells with too few cases. Be sure to save the new stratum variable using a different name. For example, if the review in step 2 above reveals that when flag1=1, v06p007a=12, and v06p007b=2 there are no cases in that cell, then you can collapse stratum 12 with the adjacent stratum 11 in a new variable called newstrat, as follows (in SPSS):

```
recode v06p007a (12=11) (else=copy) into newstrat.
```

4. Run the statistical analysis substituting the new stratum variable for v06p007a.

In the event that a very large number of cells in the matrix contain too few cases, which would require collapsing more than a few of the categories, Taylor Series may not be appropriate for the analysis. Other methods such as replication or bootstrapping may be preferable. See Lee and Forthofer (2006) for information on these methods. If cases are missing due to item nonresponse, imputation can be considered to replace the missing data. See Allison (2002) for information on imputing missing data.

Alternatives to Taylor Series: Procedures to Obtain Approximate Standard Errors

A technique to obtain standard error statistics that are approximately correct, on average, is to adjust the standard errors to reflect a study's average design effect (DEFF). The DEFF is the variance obtained with a complex sample divided by the variance that would be obtained with a simple random sample of the same size. For a study with weights scaled to a mean of 1, DEFF may be calculated as the sum of the squared weights divided by the sum of the weights. For area probability designs like the ANES, and for most complex sample surveys, complex sample designs save money at a cost in statistical power, so the DEFF is greater than 1 on average, and for most estimates.

There are two mathematically equivalent ways to adjust the standard errors to reflect the design effect. The *multiplication method* is to multiply standard errors by an adjustment factor, and the *weight adjustment method* is to apply an analogous factor to the weights. The two methods produce the same effects and the choice of one or the other is a matter of convenience.

Multiplication method. The square root of the average design effect (DEFT) is the standard error obtained with a complex sample divided by the standard error that would be obtained with a simple random sample of the same size. This means that if a study's average root design effect, DEFT, is 1.35, the design-consistent standard errors from that study are, on average, 1.35 times larger than they would be if the study had used a simple random sample. To obtain standard errors that are approximately correct, on average, simply multiply the standard error obtained when calculated by simple-random-sample methods by the square root of the study's design effect. This method can be used with standard errors from percentages, means, regression coefficients, and other statistics.

Weight adjustment method. Alternatively, the weight may be adjusted to reflect the average design effect. Calculate a new weight variable equal to the analysis weight divided by the study's average design effect, and run analyses using the new weight.[*] Let the study's design effect be represented by DEFF, the analysis weight be represented by WEIGHT, and the new weight variable be represented by DEFFWGT. The SPSS code to calculate and apply the adjustment would be as follows:

```
compute DEFFWGT = WEIGHT / DEFF .
weight by DEFFWGT.
```

You can think of this adjustment as revealing the study's effective sample size – that is, the size of the simple random sample that would have the same statistical power as the actual ANES sample.

After running the two lines of SPSS code above, subsequent commands such as regression will produce standard errors that incorporate weight adjustment.

Approximate average design effects (DEFF) and root design effects (DEFT) for selected ANES studies are shown below.

ANES Study	*DEFF*	*DEFT*
2008 Time Series post	1.57	1.25
2008 Time Series pre	1.58	1.26
2006 Pilot Study	1.82	1.35
2004 post	1.22	1.10
2004 pre	1.21	1.10
2000 post	1.26	1.12
2000 pre	1.24	1.11

[*] This description of the method assumes that the weights are scaled to a mean of 1, so they sum to the sample size. All ANES studies to date have scaled the weights this way. Many other studies scale their weights to sum to the population size, which necessitates a slightly different computation of a design effect adjustment, or a preceding step of re-scaling the weights to a mean of 1. The description here applies to ANES datasets.

1996 post	1.29	1.14
1996 pre	1.26	1.12
1992	1.14	1.07

These design effects were calculated by dividing the sum of the squared weights by the sum of the weights. (This method only works when weights are scaled to a mean of 1. For studies where weights are scaled differently, they would need to be scaled to 1 prior to this calculation.) An alternative method of calculating average design effects is to calculate design-consistent statistics for several estimates of interest using the Taylor Series method and then to take the mean of the design effects for those estimates. This alternative method may be more accurate for statistics of interest but is less general, and therefore potentially less accurate overall, than the method used in the list above.

It is better to use the Taylor series method than methods that produce more approximate standard errors, because the Taylor series method gives more accurate results. If you cannot use the Taylor series method, then you should consider replication or bootstrap methods (see for example Lee and Forthofer, 2006) or use one of the design effect adjustments just described.

These weight-adjustment procedures are sometimes called “generalized variance estimation.”

General Recommendations

Missing Data
When a respondent does not answer a survey question, a code indicates that the data are missing. For example, some variables are coded 0 to mean no, 1 to mean yes, and 9 to mean missing. For each variable in your analysis, if you want to exclude missing data, make sure that your software treats missing codes as missing data and not as numeric data. For example, if 9 is a missing code for a particular variable, be sure your software treats 9 as missing for that variable.

Data users should consider carefully the implications of missing data for their analyses. If analysts include variables with missing data in an analysis, most statistical software packages will perform listwise (also known as casewise) deletion of missing data, dropping the cases for which any variable involved in an analysis has missing values. If data are missing completely at random, listwise deletion produces unbiased estimates. When data are not missing completely at random, or when it is important to retain all cases in the analysis, other methods may be better. Some ANES data are not missing completely at random (Sherman 2000), which makes listwise deletion suboptimal. See Allison (2002) for information on handling missing data.

When choosing codes, be aware of how your software handles missing data. In SUDAAN, any categorical data with a value of 0 or less is treated as missing, so categorical data should be coded using positive integers.

Statistical Analysis and Reporting
When conducting ANES data analysis and reporting the results, the following practices are helpful.

- In any report of the results of a statistical analysis, provide enough details that someone else could replicate your analysis.
- Report the actual (unweighted) number of cases used in any analysis, not the weighted sample size. The *N* reported by most statistical software using simple-random-sample procedures with weights (including SPSS) is the weighted sample size. This figure is of little interest and generally should not be reported; it is a function of the weight scale and may be adjusted arbitrarily. Instead, report the actual number of cases used in the analysis.
- Always report unstandardized regression coefficients, whether you report standardized coefficients or not.
- Always report the scales of all variables so that readers can interpret the unstandardized regression coefficients.

- If you re-scale the variables prior to analysis, describe exactly how the transformation was calculated. Variables can be easier to compare if they are transformed to be on the same scale, such as 0 to 1.
- Always report standard errors (or confidence intervals) with every statistic. Every regression coefficient, percentage, mean, or other parameter estimate from a sample survey has an associated standard error that should be computed and reported (as in Table 1 and Table 2).
- When comparing two statistics, always report the results of an appropriate test of significance for the difference between the two statistics.

V. ANALYSIS EXAMPLES

Example Percentages and Standard Errors

Table 1 provides percentage estimates and their standard errors from the 2006 ANES Pilot Study, calculated in five different ways to illustrate the differences between the selected methods.

The "SRS, unweighted" columns present estimates and standard errors based on simple-random-sample (SRS) assumptions. Analysts should not use this method to calculate statistics with ANES datasets for which weights are provided. The estimates are presented here for comparison to the other methods. The "SRS, weighted" columns present weighted percentage estimates and standard errors based on simple-random-sample assumptions. These point estimates are correct, but the standard errors are not.

The remaining three columns present the same weighted percentage estimates – illustrating that methods of computing standard errors do not affect these estimates – with standard errors calculated in three different ways, and showing the design effects of each method in comparison to the simple-random-sample standard errors. The "Adjusted weight" column's standard errors were calculated after dividing the weight variable (V06P002) by the study's average design effect (1.82), inflating each standard error by a factor of 1.35. (These are the standard errors that result from applying the "Multiplication method" or the "Weight adjustment method" described in the previous section.) The "Design-consistent, with published strata" columns present Taylor-series estimates of standard errors. This is the method that should generally be used for analyses of the Pilot Study data whenever feasible. The "Design-consistent, with modified strata" column presents Taylor-series estimates after strata have been collapsed to permit calculation of standard errors for three sets of estimates on the second page of Table 1 that could not be calculated using the standard method because subsetting the data led to empty clusters.

The magnitude of errors that can be caused by a failure to use weights and design-consistent variance estimation is evident from Table 1. For some estimates, standard errors using simple-random-sample methods are less than half what they are when design-consistent methods are used, and unweighted point estimates are biased by several percentage points. For example, consider the estimates of the percentage of the population that is age 70 or older. The unweighted estimate (with SRS s.e.) is 11.6 (1.23), but the design-consistent estimate (weighted, with Taylor series s.e.) is 12.6 (2.87). The unweighted estimate (with SRS s.e.) for high school dropouts is 5.2 (0.85), while the design-consistent estimate is 14.4 (3.29).

Estimates in Table 1 were calculated using Intercooled Stata 9.2.

Table 1. Estimated percentages, standard errors, and root design effects from the 2006 ANES Pilot Study, calculated by alternative methods

Characteristic	SRS, unweighted		SRS, weighted		Adjusted weight			Design-consistent, with published strata			Design-consistent, with modified strata		
	Percent	s.e.	Percent	s.e.	Percent	s.e.	DEFT	Percent	s.e.	DEFT	Percent	s.e.	DEFT
Age (recode v043250)													
18-29	14.8	1.37	20.7	1.56	20.7	2.10	1.35	20.7	2.51	1.61	20.7	2.50	1.60
30-39	12.3	1.26	16.2	1.42	16.2	1.91	1.35	16.2	2.40	1.69	16.2	2.42	1.71
40-49	22.7	1.61	21.8	1.59	21.8	2.14	1.35	21.8	2.20	1.38	21.8	2.32	1.46
50-59	21.6	1.59	17.5	1.46	17.5	1.97	1.35	17.5	1.49	1.02	17.5	1.41	0.96
60-69	17.0	1.45	11.1	1.21	11.1	1.63	1.35	11.1	1.25	1.03	11.1	1.29	1.07
70 or older	11.6	1.23	12.6	1.28	12.6	1.72	1.35	12.6	2.87	2.25	12.6	2.88	2.25
Sex (v041109a)													
Male	46.1	1.92	46.8	1.92	46.8	2.59	1.35	46.8	2.45	1.28	46.8	2.45	1.28
Female	53.9	1.92	53.2	1.92	53.2	2.59	1.35	53.2	2.45	1.28	53.2	2.45	1.28
Race (recode v043299a)													
Black	10.5	1.16	11.2	1.21	11.2	1.64	1.35	11.2	1.85	1.52	11.2	1.76	1.45
Asian	2.5	0.60	2.5	0.60	2.5	0.81	1.35	2.5	0.48	0.80	2.5	0.51	0.85
Native American	0.7	0.33	0.5	0.27	0.5	0.37	1.35	0.5	0.29	1.07	0.5	0.29	1.07
Hispanic	3.7	0.73	4.2	0.77	4.2	1.04	1.35	4.2	1.16	1.50	4.2	1.11	1.44
White	78.5	1.59	76.7	1.63	76.7	2.20	1.35	76.7	1.94	1.19	76.7	1.99	1.22
Other, refused, or don't know	4.4	0.79	4.8	0.82	4.8	1.11	1.35	4.8	1.00	1.22	4.8	1.07	1.30
Education (recode V043254)													
Less than high school diploma	5.2	0.85	14.4	1.35	14.4	1.82	1.35	14.4	3.29	2.43	14.4	3.34	2.47
High school diploma/equiv.	25.3	1.68	31.4	1.79	31.4	2.41	1.35	31.4	3.13	1.75	31.4	3.19	1.79
Some college	32.9	1.81	28.5	1.74	28.5	2.34	1.35	28.5	2.05	1.18	28.5	2.11	1.21
Bachelor's degree or higher	36.6	1.86	25.6	1.68	25.6	2.27	1.35	25.6	2.30	1.37	25.6	2.21	1.32
Household income (recode V043293x)													
Under $20,000	12.7	1.28	14.9	1.37	14.9	1.85	1.35	14.9	1.97	1.44	14.9	1.96	1.43
$20,000-$39,999	18.2	1.49	18.0	1.48	18.0	1.99	1.35	18.0	1.93	1.31	18.0	1.78	1.20
$40,000-$59,999	16.9	1.44	15.1	1.38	15.1	1.86	1.35	15.1	1.71	1.24	15.1	1.59	1.15
$60,000-$79,999	15.0	1.37	16.3	1.42	16.3	1.92	1.35	16.3	1.76	1.24	16.3	1.74	1.22
$80,000-$119,999	16.1	1.42	15.8	1.40	15.8	1.89	1.35	15.8	1.61	1.15	15.8	1.46	1.04
$120,000 or more	12.4	1.27	11.0	1.20	11.0	1.62	1.35	11.0	1.59	1.32	11.0	1.36	1.13
Missing, refused, or don't know	8.6	1.08	8.8	1.09	8.8	1.47	1.35	8.8	1.22	1.12	8.8	1.22	1.12
Voter turnout 2006 (v06p775x)													
Voted	76.1	1.64	73.7	1.69	73.7	2.29	1.35	73.7	2.11	1.25	73.7	2.01	1.19
Did not vote	23.9	1.64	26.3	1.69	26.3	2.29	1.35	26.3	2.11	1.25	26.3	2.01	1.19
Voter turnout 2004 (recode v045018x)													
Voted	86.8	1.33	83.9	1.45	83.9	1.95	1.35	83.9	2.26	1.56	83.9	2.33	1.61
Did not vote	13.2	1.33	16.1	1.45	16.1	1.95	1.35	16.1	2.26	1.56	16.1	2.33	1.61
Vote in 2004 (v045026)													
John Kerry	46.6	2.13	48.4	2.13	48.4	2.87	1.35	48.4	3.97	1.87	48.4	3.83	1.80
George W. Bush	52.4	2.13	50.2	2.13	50.2	2.87	1.35	50.2	3.87	1.82	50.2	3.77	1.77
Ralph Nader	0.0	0.00	0.0	0.00	0.0	0.00	—	0.0	0.00	—	0.0	0.00	—
Other	1.1	0.44	1.4	0.50	1.4	0.67	1.35	1.4	0.80	1.60	1.4	0.80	1.60
Party ID (recode v06p680)													
Strong Democrat	25.0	1.67	26.7	1.70	26.7	2.30	1.35	26.7	2.81	1.65	26.7	2.80	1.64
Not very strong Dem.	15.4	1.39	16.2	1.42	16.2	1.91	1.35	16.2	2.00	1.41	16.2	1.95	1.38
Independent, lean Dem.	11.6	1.23	11.0	1.20	11.0	1.62	1.35	11.0	1.37	1.14	11.0	1.30	1.08
Independent	5.9	0.91	6.5	0.95	6.5	1.28	1.35	6.5	1.14	1.20	6.5	1.15	1.21
Independent, lean Rep.	8.9	1.10	8.6	1.08	8.6	1.46	1.35	8.6	1.12	1.04	8.6	1.15	1.07
Not very strong Rep.	12.3	1.26	11.7	1.24	11.7	1.67	1.35	11.7	1.06	0.86	11.7	1.07	0.86
Strong Republican	19.6	1.53	18.1	1.48	18.1	2.00	1.35	18.1	1.70	1.15	18.1	1.63	1.10
Other/apolitical/DK/refused	1.3	0.44	1.3	0.44	1.3	0.59	1.35	1.3	0.57	1.31	1.3	0.57	1.31
Bush approval (v06p790)													
Approve	28.0	1.73	27.4	1.72	27.4	2.32	1.35	27.4	1.95	1.14	27.4	1.92	1.12
Disapprove	56.7	1.91	56.8	1.91	56.8	2.57	1.35	56.8	2.51	1.32	56.8	2.53	1.33
Neither approve nor disap.	14.7	1.36	15.3	1.39	15.3	1.87	1.35	15.3	1.69	1.22	15.3	1.75	1.26
Refused	0.6	0.30	0.5	0.27	0.5	0.37	1.35	0.5	0.31	1.14	0.5	0.31	1.14

Table continues. See notes at end of table.

Table 1. Estimated percentages, standard errors, and root design effects from the 2006 ANES Pilot Study, calculated by alternative methods—*continued*

Characteristic	SRS, unweighted		SRS, weighted		Adjusted weight			Design-consistent, with published strata			Design-consistent, with modified strata		
	Percent	s.e.	Percent	s.e.	Percent	s.e.	DEFT	Percent	s.e.	DEFT	Percent	s.e.	DEFT
Trust in government (v06p656)													
Always	0.9	0.61	0.7	0.55	0.7	0.74	1.35	0.7	—	—	0.7	0.54	0.98
Most of the time	18.2	2.54	15.3	2.37	15.3	3.20	1.35	15.3	—	—	15.3	2.91	1.23
About half the time	49.4	3.30	54.9	3.27	54.9	4.42	1.35	54.9	—	—	54.9	4.79	1.46
Once in a while	27.3	2.94	24.9	2.85	24.9	3.84	1.35	24.9	—	—	24.9	3.27	1.15
Never	4.3	1.34	4.1	1.30	4.1	1.76	1.35	4.1	—	—	4.1	1.44	1.10
Internal efficacy (v06p653)													
A great deal	8.4	1.53	8.2	1.51	8.2	2.03	1.35	8.2	—	—	8.2	2.27	1.51
A lot	13.6	1.88	13.1	1.85	13.1	2.50	1.35	13.1	—	—	13.1	1.62	0.87
A moderate amount	33.1	2.59	31.5	2.55	31.5	3.44	1.35	31.5	—	—	31.5	3.53	1.38
A little	35.8	2.64	34.9	2.62	34.9	3.53	1.35	34.9	—	—	34.9	4.33	1.66
Not at all	9.0	1.58	12.3	1.80	12.3	2.43	1.35	12.3	—	—	12.3	2.91	1.61
Interest in politics (v06p630)													
Extremely interested	18.2	2.10	15.8	1.98	15.8	2.67	1.35	15.8	—	—	15.8	2.05	1.04
Very interested	39.4	2.65	38.6	2.64	38.6	3.56	1.35	38.6	—	—	38.6	5.44	2.06
Moderately interested	33.2	2.56	34.5	2.58	34.5	3.48	1.35	34.5	—	—	34.5	4.67	1.81
Slightly interested	8.5	1.52	10.1	1.63	10.1	2.20	1.35	10.1	—	—	10.1	2.35	1.44
Not interested at all	0.3	0.29	0.9	0.51	0.9	0.69	1.35	0.9	—	—	0.9	0.92	1.80
Refused	0.3	0.29	0.2	0.24	0.2	0.33	1.35	0.2	—	—	0.2	0.21	0.87
Interpersonal trust (v06p519)													
Always	1.5	0.66	1.3	0.62	1.3	0.83	1.35	1.3	0.55	0.89	1.3	0.55	0.89
Most of the time	46.9	2.71	41.7	2.68	41.7	3.61	1.35	41.7	4.26	1.59	41.7	4.31	1.61
About half the time	31.6	2.53	34.9	2.59	34.9	3.49	1.35	34.9	4.06	1.57	34.9	4.05	1.56
Once in a while	18.0	2.09	19.1	2.13	19.1	2.88	1.35	19.1	3.09	1.45	19.1	3.01	1.41
Never	1.8	0.72	2.9	0.91	2.9	1.23	1.35	2.9	1.70	1.87	2.9	1.72	1.89
Refused	0.3	0.29	0.1	0.17	0.1	0.23	1.35	0.1	0.10	0.58	0.1	0.10	0.58
Importance of religion (v06p552)													
Important	73.9	1.69	73.8	1.69	73.8	2.28	1.35	73.8	2.69	1.59	73.8	2.72	1.61
Not important	25.9	1.69	26.1	1.69	26.1	2.28	1.35	26.1	2.69	1.59	26.1	2.71	1.60
Refused	0.1	0.15	0.1	0.12	0.1	0.16	1.35	0.1	0.09	0.74	0.1	0.09	0.74

Notes: — Statistic cannot be calculated by standard techniques.
"SRS, unweighted" estimates are calculated assuming simple random sampling, without weights. This method is not recommended.
"SRS, weighted" estimates are calculated assuming simple random sampling, weighted by v06p002. This method of calculating standard
"Adjusted weight" estimates are calculated with v06p002 adjusted by the average design effect (v06p002/1.82), reducing the effective
"Design consistent, with published strata" estimates are weighted by v06p002 and use Taylor series standard errors based on stratum and
"Design consistent, with modified strata" estimates are weighted by v06p002 and use Taylor series standard errors based on collapsed
s.e. is standard error, also known as sampling error.
DEFT is the square root of the design effect (DEFF). Each DEFT equals the s.e. in the adjacent column divided by the s.e. in the "SRS,
Source: 2004 National Election Study and 2006 ANES Pilot Study (datasets).

Example of Regression Analysis

Table 2 presents logistic regression analyses with the 2008 ANES Time Series data, calculated three different ways to illustrate the differences between the selected methods.

In each case the same model is presented. The purpose of the model is only to illustrate an analysis, so the choice of variables was arbitrary. The dependent variable is voting for Barack Obama in the 2008 general election. It is coded 1 if the respondent voted for Obama and 0 if the respondent voted for another candidate. Nonvoters were excluded from the analysis. The independent variables are party ID, feeling thermometers for Obama and McCain, the respondent's sex and educational attainment, whether the respondent is black, belief about the Bible being the word of God, whether the Iraq war was worth the cost, whether homosexuals should be allowed to serve in the armed forces, and the interviewer's assessment of whether the respondent seemed well informed. Each independent variable was re-scaled to range from 0 to 1. Exhibit 2 shows the Stata code for the analyses.

Table 2. Logistic regression models of vote choice for Obama, 2008 ANES Time Series Study

	Model 1			Model 2			Model 3		
	Unweighted, SRS SEs			Weighted, SRS SEs			Weighted, design-based SEs		
Independent variable	Coef.	s.e.	p	Coef.	s.e.	p	Coef.	s.e.	p
Party ID	-3.4 ***	0.47	0.000	-3.7 ***	0.55	0.000	-3.7 ***	0.39	0.000
Obama FT	7.7 ***	0.72	0.000	8.2 ***	1.00	0.000	8.2 ***	1.31	0.000
McCain FT	-5.4 ***	0.74	0.000	-5.4 ***	0.96	0.000	-5.4 ***	0.77	0.000
Bible word of God	-1.1 **	0.40	0.006	-1.2 **	0.46	0.009	-1.2 **	0.55	0.030
Gays in military	**0.6**	**0.40**	**0.132**	**0.9**	**0.51**	**0.092**	**0.9 ***	**0.37**	**0.024**
Iraq war worth costs	**-0.8 ***	**0.31**	**0.012**	**-0.7**	**0.38**	**0.062**	**-0.7**	**0.36**	**0.057**
Appeared informed	**1.1 ***	**0.52**	**0.032**	**1.1**	**0.72**	**0.144**	**1.1**	**0.66**	**0.115**
Education	-3.7 ***	1.05	0.000	-4.2 **	1.32	0.001	-4.2 **	1.19	0.001
Female	0.3	0.27	0.277	0.1	0.32	0.872	0.1	0.25	0.833
Black	2.5 ***	0.63	0.000	2.6 ***	0.76	0.000	2.6 ***	0.68	0.000
Constant	2.7 *	1.05	0.011	2.9 *	1.34	0.030	2.9 **	1.01	0.005

*** p<.001; ** p < .01; * p < .05

Notes: Model 3 employs the recommended method. Models 1 and 2 use methods that are not recommended. Coef. is the logistic regression coefficient. s.e. is standard error. p is the p-value, the probability of the coefficient differing from zero due to sampling error by as much as the observed value in the event that the true population coefficient is zero. n = 1,424 in each model.

Source: 2008 ANES Time Series study (dataset).

Three versions of the logit model are presented in Table 2. *Model 1* is an example of an incorrect analysis: ordinary procedures that assume simple random sampling are used, and the analysis is unweighted. In Model 1, the regression coefficients as well as the standard errors are incorrect. *Model 2* is an example of a better but still flawed analysis: the data are weighted, but the standard errors are not adjusted to account for the survey design. In this model the regression coefficients are correct but the standard errors are not. *Model 3* is correct. It is weighted and uses the Taylor Series adjustments to compute design-consistent standard errors.

The differences between these models illustrate some of the differences between design-consistent methods and other approaches. In some important respects, the three models are substantively the same. All three show that party identification and candidate feeling thermometers are strong and statistically significant predictors of the vote. However, the recommended analysis (model 3) supports at least three substantively different conclusions than the unweighted model (model 1). The rows showing these differences are set in bold in Table 2.

In the unweighted model, preferences regarding gays in the military are not significant. In the design-based model, the effect size for this variable is larger by half and the standard error is smaller, and the variable is statistically significant at the p<.05 level. Although design-based standard errors in ANES data are, on average, larger than standard errors would be with a simple random sample of the same size, it is not unusual for the design-based standard errors to be smaller for some variables, and this is an example of that.

The next two rows of Table 2 show coefficients for whether the respondent thought the Iraq war was worth the cost and whether the interviewer thought the respondent seemed well informed. Both variables are statistically significant at the p<.05 level in the unweighted analysis, but in the design-based analysis, neither meets this threshold of significance.

This example shows that analyzing ANES data without weights and without design-based corrections to standard errors may lead to misleading conclusions. Moreover, it may lead to both type-1 and type-2 errors.

The code to replicate these analyses in Stata is shown in Exhibit 2.

Exhibit 2. Stata program (.do file) for regression analyses in Table 2

```
* Code for Table 2
* allocate 50MB of memory to accommodate the large data file
set mem 50M
*open data containing stratum and PSU as V081205 and V081206
use "C:\Documents and Settings\Matthew DeBell\My Documents\ANES\how to analyze\2008TS w
stratum.dta"
*set missing values
recode V083097 V083098a V083098b V083037a V083037b V081101 V085044a V083184 V083217 V083216x
V083184 V083217 V083103 V083303 V083212x V081102 (-9/-1=.)
* variable setup; set missing values and generate variables
gen voteobama=.
replace voteobama=1 if V085044a==1
replace voteobama=0 if (V085044a==3 | V085044a==7)
* generate the subpopulation identifier variable
gen subp=0
replace subp=1 if (voteobama==0 | voteobama==1)
gen married =0
replace married =1 if V083216x==1
gen pid=.
replace pid=1 if (V083097==1 & V083098a==1)
replace pid=2 if (V083097==1 & V083098a==5)
replace pid=3 if V083098b==5
replace pid=4 if V083098b==3
replace pid=5 if V083098b==1
replace pid=6 if (V083097==2 & V083098a==5)
replace pid=7 if (V083097==2 & V083098a==1)
gen black=.
replace black=1 if (V081102==2 | V081102==6 | V081102==7)
replace black=0 if (V081102==1 | V081102==4 | V081102==5)
replace V083184 =. if V083184 ==7
* create intelligible variable names
rename V083037a obamaft
rename V083037b mccainft
rename V081101 sex
rename V083184 bibleWOG
rename V083217 edu
rename V083103 iraqworth
rename V083303 obsinfo
rename V083212x gaymil
* scale all the variables 0-1
replace pid = (pid-1)/6
replace obamaft = obamaft/100
replace mccainft = mccainft/100
generate female = sex-1
recode bibleWOG (1=1) (2=.5) (3=0)
replace edu = edu/17
recode iraqworth (1=1) (5=0)
recode gaymil (1=1) (2=.667) (4=.333) (5=0)
recode obsinfo (1=1) (2=.75) (3=.5) (4=.25) (5=0)
*set up the survey procedures
svyset [pweight=V080102], strata(V081206) psu(V081205)
* regressions: unweighted SRS, weighted with SRS SEs, and weighted with Taylor Series SEs
logit voteobama pid obamaft mccainft female bibleWOG edu iraqworth black gaymil obsinfo
logit voteobama pid obamaft mccainft female bibleWOG edu iraqworth black gaymil obsinfo
[pweight=V080102]
svy, subpop(subp): logit voteobama pid obamaft mccainft female bibleWOG edu iraqworth black
gaymil obsinfo
```

REFERENCES

Allison, P.D. (2002). *Missing Data*. Quantitative Applications in the Social Sciences series, vol. 136. Thousand Oaks, CA: Sage.

Crawley, M.J. (2005). *Statistics: An Introduction Using R*. New York: John Wiley.

Crawley, M.J. (2007). *The R Book*. New York: John Wiley.

DeBell, M., and Krosnick, J.A. (2009). *Computing Weights for American National Election Study Survey Data*. ANES Technical Report Series, no. nes012427. Available online at ftp://ftp.electionstudies.org/ftp/nes/bibliography/documents/nes012427.pdf

Lee, E.S., and Forthofer, R.N. (2006). *Analyzing Complex Survey Data*. Second edition. Quantitative Applications in the Social Sciences series, vol. 71. Thousand Oaks, CA: Sage.

Levy, P.S., and Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*. 3rd ed. New York: Wiley.

Sherman, R.P. (2000). "Tests of Certain Types of Ignorable Nonresponse in Surveys Subject to Item Nonresponse or Attrition." *American Journal of Political Science*, *44*, 356-368.