Computing Weights for American National Election Study Survey Data

Matthew DeBell

Jon A. Krosnick

Stanford University

September 1, 2009

This work was supported by National Science Foundation Grants SES-0535332 and SES-0535334 (PIs: Jon A. Krosnick and Arthur Lupia), Stanford University, and the University of Michigan. We thank Walter Mebane and Vince Iannacchione for their advice. Opinions, findings, conclusions, and recommendations expressed here do not necessarily reflect the views of the National Science Foundation, Stanford University, or the University of Michigan. Contact: debell@stanford.edu, krosnick@stanford.edu, or anes@electionstudies.org.

Suggested Citation
DeBell, Matthew, and Jon A. Krosnick. 2009. *Computing Weights for American National Election Study Survey Data*. ANES Technical Report series, no. nes012427. Ann Arbor, MI, and Palo Alto, CA: American National Election Studies. Available at http://www.electionstidies.org

Abstract

This report describes methods to calculate weights for ANES studies that account for the sampling design and match population benchmark for selected variables. ANES data are based on complex sample designs and must be weighted to adjust for the sample design in order to generalize to the population. Weights for such analysis must adjust for unequal probability of household selection and for respondent selection within households. Additionally, the method described here includes nonresponse adjustment and post-stratification raking when needed to match known population benchmarks. Poststratification factors should be selected based on a comparison to benchmark statistics, to correct notable differences from benchmarks. The selection of poststratification factors, coding categories, and trimming thresholds should be made to strike a balance between variance in the weights and accuracy of the estimates. In panel studies, weights should be calculated for each combination of waves that will be analyzed together. Panel weights may also be developed to adjust for panel attrition. Details are described at each step in the process.

Introduction

This memo reports the results of an investigation of the weight variables provided with American National Election Study (ANES) survey datasets. The memo is based upon advice provided by an advisory committee of five experts in survey methodology and statistics:

Douglas Rivers (chair)
Professor of Political Science; Senior Fellow, Hoover Institution; Research Fellow, Stanford Institute for the Quantitative Study of Society
Stanford University

Martin Frankel
Professor of Statistics and Computer Information Systems
Baruch College, City University of New York

Colm O'Muircheartaigh
Senior Fellow, National Opinion Research Center
Professor of Public Policy and Statistics
University of Chicago

Charles Franklin
Professor of Political Science
University of Wisconsin, Madison

Andrew Gelman
Professor of Statistics and Political Science
Columbia University

Based upon the committee's advice, general procedures for constructing weights for ANES studies have been specified and are described in this document.[1]

Philosophy of Weighting

There is no single correct way to construct weights for a particular study. Weighting, and poststratification weighting in particular, involves making judgments about the relative value of

[1] Although the recommendations in this memo were all endorsed by committee members, not every committee member agrees with every recommendation in the memo.

accuracy and low variance and about the variables for which accuracy is most desirable. If constructed well, weights normally enhance accuracy at the price of increased variance (which means less precise estimation and fewer statistically significant findings), and for any study there are many reasonable ways to create weights based on different reasonable choices about these tradeoffs.

The range of potential reasonable choices about weighting can increase when data are collected over time. For many of the kinds of studies that ANES produces, such as its Time Series and Panel Studies, weighting also requires assumptions about how key sample properties change over time. In particular, different points of view about the causes and unseen consequences of panel attrition or changes in unit non-response over time can lead different scholars to advocate different weighting methods.

This memo describes one reasonable general approach to weighting ANES studies, though other approaches may be equivalently legitimate. In some cases, our approach leads us to provide multiple weights for a single data collection – a strategy that allows ANES to serve diverse analytic needs.

Single Cross-Sectional Face-to-Face Study[2]

To build weights for a single cross-sectional face-to-face survey involving in-home interviewing and a multi-stage area probability sample design (involving first selection of PSUs, then area segments, then housing units, and then respondents within households), the following steps should be implemented.

Unequal Probability of Household Selection

1. Weights must be constructs to adjust for unequal probabilities of household selection into the sample. If stratified sampling was implemented, or if over-sampling of particular types of households was done (such as of households in particular geographic areas), then weighting must correct for these by-design inequalities. The greater the household's probability of inclusion in the study's sample, the smaller that household's weight should be. This probability is the product of the probability of selecting each unit at each selection stage; for example, in a sample with PSUs, area segments, and households within area segments, the probability of household selection is the product of the probability of the PSU selection, the area segment selection within the PSU, and the

[2] The steps for an RDD telephone study are the same, except that step 1 must be revised to account for the RDD sample design.

household selection within the area segment. The weight for the household's probability of selection is equal to the inverse of the household's probability of selection.

Unequal Probability of Respondent Selection within Households

2. Weights must be constructs to correct for unequal probability of selection of respondents within households. If one respondent to be interviewed was selected randomly within each household, then the number of residents of the household who were eligible to be interviewed in the study must be used to for this purpose.

Unequal Rates of Nonresponse

3. Nonresponse adjustments are based on known differences in response rates (RRs) among sample elements with specific characteristics. Typically, the respondent's Census region, cluster, metropolitan status, or other geographic characteristics are the only characteristics known about all cases that did not complete an interview. Look at the response rates by region or another sample group in which the RR is known and which may be associated with variables of interest in the study. If regions or groups differ in their response rates, then a nonresponse adjustment may be worth performing. When the RRs are unequal, include a weighting factor proportional to the inverse of the RR in the respondent's group.

Post-stratification Using Known Population Parameters

4. Conduct a benchmark comparison.

   a) Using the study questionnaire, identify a set of variables likely to be measured with little error in the survey and with a low item nonresponse rate and compare their distributions with "truth" benchmarks from a reliable source such as the Current Population Survey (CPS). When making these comparisons, focus on variables measured identically (or as comparably as possible) in the survey and the CPS. Whenever possible, such variables should include age, sex, race/ethnicity, educational attainment, Census region, metropolitan/non-metropolitan status (or urban/suburban/rural status),[3] marital status, home tenure, household size, and

[3] Metropolitan status can be a problematic benchmark variable because it may not be available for the respondents and because accurate benchmark statistics may be difficult to obtain. CPS public-use datasets have a proportion of metropolitan status data set to missing to protect respondent confidentiality, calling into question the accuracy of benchmarks calculated with this edited variable. Use metropolitan status as a poststratification factor only if accurate benchmark statistics can be obtained using a complete dataset.

perhaps others such as student status, if available, or income, if measured accurately with a low item nonresponse rate.[4]

b) For post-election surveys in applicable years, include presidential vote choice in the comparison.

c) For post-election election surveys, include voter turnout in the comparison. The benchmark for voter turnout should be the percentage of the voting-eligible population that cast a ballot in the presidential election. (It should not be the percentage of the voting-*age* population, which is often more widely reported.) Such turnout estimates are reported by the United States Elections Project at elections.gmu.edu/voter_turnout.htm

d) ANES statistics calculated for initial benchmark comparisons should use an ANES weight that accounts for unequal probabilities of household selection, respondent selection, and nonresponse, as described above in items 1–3.

5. Interpret the benchmark comparison.

   a) If no notable demographic discrepancies are observed, then post-stratification weighting with demographics may not be necessary.

   b) If notable demographic discrepancies are observed for some variables, post-stratification weighting with these demographics should be done.

   c) What constitutes a notable demographic discrepancy is a matter of judgment. Often, demographic discrepancies exceeding 5 percentage points are "notable" and discrepancies less than 2 percentage points are not. Discrepancies in the 2 to 5 point range may be notable if the characteristic is of special interest for the study or is strongly associated with key outcome variables such as voter turnout or candidate choice.

6. Select and prepare variables for poststratification.

   a) Select demographic variables for poststratification that have notable differences from benchmarks.

   b) Select only poststratification factors that are believed to be measured accurately and that have low item nonresponse rates. Variables with item nonresponse rates exceeding 5 percent (as income often does) may not be suitable poststratification factors.

[4] Survey questions about income often have high item nonresponse rates. Income nonresponse is likely to be nonrandom. Do not use income as a poststratification variable if the item response rate is lower than about 95 percent.

c) In addition to demographic characteristics, voter turnout rate and the distribution of votes for president reported by the federal government may be poststratification factors for *supplementary* versions of the weights if ANES estimates have notable differences from benchmarks.

d) Comparisons of estimates over time are central to the ANES mission. Therefore, to the extent that different studies have the same design and response rate, and are assumed to have identical nonresponse characteristics, weighting procedures (including the selection of poststratification factors) should be as similar as possible so that comparisons over time will not be affected by weights. However, we do not assume that nonresponse characteristics are constant over time.

e) When sampling designs, response rates, nonresponse characteristics, and research questions differ between studies, different weighting procedures may be warranted to maximize data quality. Weighting procedures (including the selection of poststratification factors) may therefore differ from study to study. (To the extent that nonresponse characteristics are known, it is not typical for them to remain constant from study to study over the years. Therefore, using the same weighting method across time only yields the appearance of consistency when nonresponse varies, and weights should be tailored to specific studies.)

f) Post-stratification weighting may be done on two-way marginals (e.g. sex × age) or sets of two-way marginals (e.g. sex × age and sex × race) as well as one-way marginals. A decision to use two-way marginals should be based on a benchmark report that shows discrepancies on those marginals that are not corrected by raking to one-way marginals, or on a desire to optimize the weights for the analysis of a population subgroup (see item 12, below).

g) Before implementing poststratification, employ a simple imputation procedure to replace missing values of demographic variables to be used in the raking only if such a procedure can be implemented readily enough to justify the benefit, which is likely to be slight. If imputation is not implemented, include cases with missing data on raking factors in the dataset and assign them a final weight equal to their nonresponse-adjusted base weight.

h) If any variables chosen as poststratification factors are continuous or have more than approximately 6 categories or have a category that contains less than approximately 5 percent of the sample, recode each such variable into a categorical variable with no more than 6 categories, none of which contain less than 5 percent of the sample. For example, a race/ethnicity variable with the categories white, black, Hispanic, and other should not be used if “other” contains only 4 percent of the sample. Instead, “other” should be combined with another category.

7. Post-stratification weighting should be implemented using raking and not cell weighting, because raking is more flexible in the use of more demographic variables at once.

8. Raking can be accomplished by writing code from scratch in any powerful statistical software or by using built-in tools in Stata or WESVAR, the SURVWGT module for Stata written by Nick Winter, programs in R such as the Survey package by Thomas Lumley (see http://faculty.washington.edu/tlumley/survey/), or with the "RAKING" SAS macro developed by Izrael, Hoglin, and Battaglia, at www.abtassociates.com/attachments/**sas**balancingweighted.pdf

9. Rake.

   a) Multiply the nonresponse-adjusted base weight for each case by the factor required to match the target population percentage for the first (arbitrarily ordered) poststratification factor.[5]
   b) Multiply the product of step *a* by the factor required to match the target population percentage for the next (arbitrarily ordered) poststratification factor.
   c) Continue for each of the remaining poststratification factors.
   d) At each step, cap (truncate) extreme weights, if there are any. Limit the range of weights to a maximum of about 5 times the mean weight (i.e., 5 if the mean weight is 1.0) by recoding any weight greater than 5x to 5x. Record the original (uncapped) values for later review and to permit documentation of the extent of capping. (Later steps may warrant an amendment to the truncation threshold of 5.) Truncate large values but do not truncate small values (near 0) because large weight values increase the potential for outliers to affect analyses and are more likely to inflate variance, while small values do not have these consequences.
   e) The factors applied to make the estimates match the population on later factors typically cause the estimates for the earlier factors to diverge from the targets. To fix this, repeat the entire process, and continue repeating it until all the estimates converge on the benchmarks or until the current iteration produces no change from the previous iteration. This may require several repetitions.

10. After raking, modifications and re-raking are often warranted. Evaluate the raked weights, revise the approach, rake again, and repeat as necessary.

[5] For example, if the sample is 60 percent female and 40 percent male (when weighted using the base weight) while the population is 52 percent female and 48 percent male, then the base weight for females would be multiplied by .52/.60 = .87 and the base weight for males would be multiplied by .48/.40 = 1.2. The percentages for sex will match the population when weighted by the product of this adjustment.

a) Examine the cases that were capped at item 9d. If most share a specific characteristic (such as the respondents being of a specific age group, racial group, sex, or sharing another characteristic), then capping weights is likely to cause biased estimates on that characteristic. Examine the effects of the cap on this group, and if the cap changes estimates for this group, raise the cap to 6, 7, or 8, using the lowest cap that minimizes or eliminates the cap's effect on the estimates for the group.

b) After raking, conduct a new benchmark comparison (item 4 above) using the new raked weights.

c) If the benchmark comparison shows that the survey estimates differ from the benchmarks that were used as raking factors, consider increasing weight caps (imposed at item 9d) to 6, 7, or 8.

d) Examine the effects of raking on variables not used as raking factors. If any of these estimates show a greater difference from the benchmarks using the new weights, try raking with a revised poststratification approach (item 10g).

e) Examine the coefficient of variation using the new weights, and the design effect using the new weights. If these are greatly inflated (e.g., if the design effect with the new raked weights exceeds the design effect prior to raking by more than 0.5), try raking with a revised poststratification approach to minimize this effect.

f) Examine the coefficient of variation for the full sample as well as for subsets of interest, such as members of minority groups if the study is intended to allow analyses of these groups.

g) If revising the poststratification approach: One way to limit the coefficient of variation is to collapse categories in the variables used for raking. Another is to eliminate or replace poststratification factors. A third is to adjust the cap. Increasing the cap can make benchmark estimates more accurate.

Final Weight Design

11. Scale the weights so that they average 1.00000.

Weights Suitable for Subgroups

12. Weights that match benchmarks for the full sample will not necessarily match benchmarks when the data are subset to a specific group, such as voters or members of a specific race/ethnicity group. When subgroup analysis is a study objective, it may be desirable to compute weights tailored to representation of the subgroup. One way to make weights suitable for analysis with the full sample or with subgroups is to rake on two-way marginals. For example, raking on sex × age would be expected to promote representativeness of the age distribution when the data were subset to one sex. Another

approach to making weights suitable for the analysis of subgroups is to drop cases outside the group, run benchmark statistics for that subgroup, and create poststratified weights for the subgroup only.

Supplemental Weights for Accuracy

13. If constructed well, weights normally enhance accuracy at the price of increased variance (which means less precise estimation and fewer statistically significant findings), and for any study there are many reasonable ways to create weights based on different reasonable choices about these tradeoffs. The procedures described above aim for an accuracy/variance tradeoff that emphasizes relatively low variance, for which the price may be low accuracy for some estimates. When the study design calls for maximizing accuracy on specific demographic factors, or when the procedures described above only yield weighted estimates with many substantial differences from the benchmarks, develop a *supplementary* weight (or weights) using procedures that allow the design effect due to weighting to increase by more than 0.5 (see item 10e). Do so by allowing more categories in the poststratification variables and/or allowing smaller categories in those variables (see section I, item 6h), using a higher cap or omitting the capping step (item 9d), adding or replacing poststratification factors, or a combination of these revised approaches.

Dataset and Documentation

14. In the dataset, provide separate variables to indicate all levels of stratification and clustering, such as stratum, PSU/cluster, and area segment. These can be indicated by assigning each unit an arbitrary number without specifying exactly which geographic location is indicated by each number.

15. In the dataset, include the base weights created prior to poststratification.

16. In the dataset or documentation, describe the components used at each step to create the analysis (poststratified) weight.

17. If multiple versions of analysis weights are included in a dataset, write clear documentation so that novice users can easily find the default weight that is most appropriate for them to use.

18. In the documentation, describe precisely the characteristics and size of the target population and all of the benchmarks used for raking. Describe the inflation factor by

which weights should be multiplied if analysts wish to make the weights sum to the population size.

19. In the documentation, list the variables on which raking was performed and explain that sampling errors on percentages and on means from these variables are not meaningful due to raking.

20. In the documentation, report design effects for selected statistics of interest.

21. In the documentation, if weights were truncated, describe exactly how.

Two-Wave Panel

All of our presidential-year surveys involve two-wave panels based on area probability samples. (In some years, additional cases were interviewed by telephone via RDD.) With datasets based on area probability samples, researchers often want to analyze the subset of respondents who were interviewed both pre-election and post-election. Such surveys have often involved systematic attrition, such that people who were minimally interested in politics were especially likely not to be interviewed post-election.

Since we measured interest in politics (and lots of other variables) in our pre-election interviews, we can build weights that correct for systematic attrition using the pre-election measurements of interest in politics and/or other variables. Based on the assumption that parameter estimates of interest in politics are more accurate in the pre-election wave than in the (attrition-diminished) post-election wave, weights that correct for attrition using pre-election measurements should increase the accuracy of other estimates.

Therefore, to construct weights in a two-wave panel, we would follow this procedure.

1. Weight the first wave using the procedures described in section I above.

2. If notable discrepancies are observed between people who were and were not interviewed post-election in terms of some variables measured in the first wave (such as partisanship and interest in politics), post-stratification weighting of the second wave with these variables should be done if the attrited cases are similar to the retained cases in terms of other pre-election variables (since this procedure is based on the assumption that, e.g., people with little interest in politics who drop out are the same as people with little interest in politics who remain in the study). To determine this, compute cross-tabulations comparing the attrited and retained cases in terms of the distributions of many

pre-election variables. If notable discrepancies are observed attrition weighting should not be performed. Guidelines for what constitutes a notable difference were presented at Item I-5-c.

3. To calculate the analysis weight for the second wave, rake using the following procedure:

    a) Multiply the first-wave analysis weight for each case by the factor required to match the target population percentage for the first (arbitrarily ordered) poststratification factor.
    b) Using the product of the previous step, repeat the adjustment for each of the remaining poststratification factors and each of the attrition factors, if any, identified in step 2 above.
    c) At each step, truncate extreme weights as they were truncated on the weights for the first wave.
    d) Continue re-applying the adjustment factors until all the estimates converge on the benchmarks or until the current iteration produces no change from the previous iteration. This may require several repetitions.

4. Follow the evaluation and revision procedures described for a cross-sectional study in part I, step 10 .

5. Release at least two analysis weights for a two-wave panel: (1) one weight for analysis of the first wave alone, and (2) a second weight for analysis of the second wave alone or in conjunction with the first wave. (Note that in our two-wave panels, all wave-2 respondents responded to the first wave, so the set of respondents who responded to wave 2 is identical to the set of respondents who responded to both waves 1 and 2.)

Multi-Wave Panel

Weighting for multi-wave panel study is an extension of weighting for a two-wave study.

In multi-wave panels, respondents to wave *W* may or may not have completed wave *W*-1. As a result, a unique subset of the full sample may have completed any given wave or any given set of waves.

The optimal weight for any given data analysis in a multi-wave study would be calculated using the subset of respondents included in that analysis. For example, an analysis of responses to questions from wave 1 would optimally use a weight calculated using the respondents who completed the wave 1 survey; an analysis of responses to questions from waves 2, 5, and 8 would be weighted using the subset of respondents who completed wave 2 and wave 5 and wave 8. This approach is optimal because it tailors the weight to the data being analyzed, assuring that missing data due to unit nonresponse (i.e., nonresponse to an entire wave) do not bias the weighted estimates. It also assures that questionnaire data that could be used for the analysis are not excluded due to weights being set to zero. However, in a multi-wave study, it may be impractical to create weights that are optimized for every possible analysis, because the number of combinations of waves is very large. For example, a 12-wave panel study has 4,095 possible combinations of one or more waves.

To construct weights in a multi-wave panel, follow this procedure.

1. Calculate the total possible number of combinations of waves that could be analyzed.[6] If this number is reasonably small (perhaps <30), calculate a weight optimized for each wave and each combination of waves by using the procedures described above for a two-wave study. If this number is impractically large (perhaps 30 or more), follow steps 2 through 5 below.

2. Release "cross-sectional weights" that are designed for the analysis of each respective wave individually. There would be a weight optimized for analysis of data from wave 1 only, and one optimized for analysis of wave 2 only, etc.

3. Release a "cumulative weight" that is designed for the analysis of all waves at the same time. This weight would be non-zero for respondents who completed every wave of the panel study and would be zero for all other respondents.

4. Consider releasing additional cumulative weights that are optimized for the analysis of a given wave and all prior waves. This weight would be non-zero for respondents who completed every wave of the panel study through wave *W* and would be zero for all other respondents. A five-wave study would have cumulative weights for waves 2, 3, 4, and 5. One weight would be for respondents who completed waves 1 and 2; one would be for respondents who completed waves 1, 2, and 3; one would be for respondents to waves 1, 2, 3, and 4; and one would be the cumulative weight for respondents to all 5 waves.

5. Consider releasing additional weights tailored to specific analyses of interest. For example, in an 8-wave study, if the study design calls for an analysis of waves 2, 5, and 8 together, consider developing a weight optimized for this analysis. Such a weight would

[6] The number is given by $n! / \{k!(n - k)!\}$, where n is the number of waves in the panel study and the calculation is repeated for values of $k = 1$ through n and the results of each are summed. More tersely, the calculation is $2^n - 1$. A five-wave study has 31 possible combinations. Adding a wave doubles and adds 1 to the number of combinations.

be non-zero for each respondent who completed wave 2 and wave 5 and wave 8 and would be zero for other respondents.

Documentation

Much of the ANES user community is either unaware or unconvinced that ANES data should be analyzed with weights and design-consistent significance tests. To educate and help users, the following should be provided in documentation:

1. ANES will publish a "How to Analyze ANES Data" document, featured prominently on our web site and referenced in other study documentation, that discusses the use of weights and design-consistent estimates and provides examples of how to produce weighted estimates with design-consistent standard errors using one or more mainstream statistical software packages.

2. We will make prominent, explicit statements in the documentation of all future studies that all analysis of ANES data intended to generalize to the population must be weighted. If methodologists wish to incorporate design factors into models, the study design may be accounted for that way instead, but properly specifying the design in a model may be difficult. We will tell users that unweighted percentages and regression coefficients do not constitute legitimate estimates of population parameters.

3. We will note that unweighted analysis is legitimate to assess the findings of a survey experiment within the participating sample. However, the results of unweighted analyses cannot be generalized to estimate the effects that would be observed in the population or in weighted analysis.

4. We will report each study's design effect for the full sample and for subsets of the sample likely to be analyzed separately.