Pew Research Center

FOR RELEASE JANUARY 26, 2018

For Weighting Online Opt-In Samples, What Matters Most?

The right variables make a big difference for accuracy. Complex statistical methods, not so much

BY *Andrew Mercer, Arnold Lau, and Courtney Kennedy*

FOR MEDIA OR OTHER INQUIRIES:

Andrew Mercer, Senior Research Methodologist
Courtney Kennedy, Director, Survey Research
Hannah Klein, Communications Associate

202.419.4372

www.pewresearch.org

RECOMMENDED CITATION

Pew Research Center, January, 2018, "For Weighting Online Opt-In Samples, What Matters Most?"

About Pew Research Center

Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping America and the world. It does not take policy positions. It conducts public opinion polling, demographic research, content analysis and other data-driven social science research. The Center studies U.S. politics and policy; journalism and media; internet, science and technology; religion and public life; Hispanic trends; global attitudes and trends; and U.S. social and demographic trends. All of the Center's reports are available at www.pewresearch.org. Pew Research Center is a subsidiary of The Pew Charitable Trusts, its primary funder.

For Weighting Online Opt-In Samples, What Matters Most?

The right variables make a big difference for accuracy. Complex statistical methods, not so much

A growing share of polling is conducted with online opt-in samples.[1] This trend has raised some concern within the industry because, while low participation rates pose a challenge for all surveys, the online opt-in variety face additional hurdles. By definition they do not cover the more than 10% of Americans who don't use the internet. The fact that potential respondents are self-selected means that there is still substantial risk that these samples will not resemble the larger population. To compensate for these challenges, researchers have employed a variety of statistical techniques, such as raking, propensity weighting and matching, to adjust samples so that they more closely match the population on a chosen set of dimensions. Researchers working with online opt-in samples must make a great many decisions when it comes to weighting. What factors should guide these decisions, and which ones are most consequential for data quality?

A new Pew Research Center study adds to the survey field's broader efforts to shed light on these questions. The study was based on over 30,000 online opt-in panel interviews conducted in June and July of 2016, with three vendors, and focuses on national (as opposed to state or local level) estimates. We evaluated three different weighting techniques, raking, propensity weighting and matching, both on their own and in combination. Every method was applied using two sets of adjustment variables: basic demographics (age, sex, race and ethnicity, education, and geographic region), and a more extensive set that included both demographics and a set of variables associated with political attitudes and engagement (voter registration, political party affiliation, ideology and identification as an evangelical Christian). Each procedure was performed on simulated samples ranging in size from n=2,000 to n=8,000.

The procedures were primarily appraised according to how well they reduced bias on estimates from 24 benchmark questions drawn from high-quality federal surveys.[2] They were also compared in terms of the variability of weighted estimates, accuracy among demographic subgroups, and their effect on a number of attitudinal measures of public opinion.

---

[1] Online opt-in samples are comprised of people who joined a survey panel or completed a one-off survey while using the internet.

[2] Survey estimates that are closer to the population benchmark values (from sources like the U.S. Census Bureau's American Community Survey) are taken to be more accurate than those farther away from the benchmarks. Using benchmarks is a useful though imperfect approach for estimating bias. For a discussion of its limitations, see the bottom of Chapter 1 of Pew Research Center's 2016 report "Evaluating Online Nonprobability Surveys."

Among the key findings:

- **Even the most effective adjustment procedures were unable to remove most of the bias.** The study tested a variety of elaborate weighting adjustments to online opt-in surveys with sample sizes as large as 8,000 interviews. Across all of these scenarios, none of the evaluated procedures reduced the average estimated bias across 24 benchmarks below 6 percentage points – down from 8.4 points unweighted. This means that even the most effective adjustment strategy was only able to remove about 30% of the original bias.

How did Pew Research Center evaluate different adjustment procedures for online opt-in samples?

1. Collected data from three opt-in panels

We fielded three large surveys with different online opt-in panel vendors, each with roughly 10,000 interviews. Each used the same questionnaire, which contained a combination of federal benchmarks, demographics, and other attitudinal and behavioral measures.

2. Simulated surveys of different sample sizes

Separately for each of the three surveys, we took random subsamples with sizes from 2,000 to 8,000 in increments of 500. This let us test how well different procedures work for a range of sample sizes.

3. Compared adjustment methods and variables

For each subsample, we tested several adjustment methods. Each was applied using two sets of variables: only demographics, and demographics + political variables. We then compared the accuracy of survey estimates produced by each adjustment.

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

- **When it comes to accuracy, choosing the right variables for weighting is more important than choosing the right statistical method.**[3] Adding a set of politically focused variables to the weighting adjustment reduced the average estimated bias by an additional 1.4 percentage points relative to adjusting only on basic demographics (e.g., age, education, race). While that might seem small, a difference of 1.4 points in the average implies that about 36 percentage points of bias were removed overall, but spread out across all 24 variables. Benchmarks most strongly associated with the political adjustment variables saw the largest improvements. In contrast, the use of more complex statistical methods never reduced the average estimated bias by than 0.3 points beyond what was achieved with raking, the most basic statistical method evaluated.[4]

- **The benefits of adding political variables to adjustment differ by survey topic.** Perhaps not surprisingly, benchmarks related to political engagement saw the largest improvement with the addition of political adjustment variables. Unweighted, these benchmarks had an average estimated bias of 22.3 percentage points, more than any other topic. While demographic weighting reduced the average bias by an average of 2.9 points, the effect of adding political adjustment variables was four times as large, reducing bias by 11.7 points and cutting the average estimated bias nearly in half (to 10.6 percentage points). Benchmarks pertaining to civic engagement and technology use also benefited disproportionately from political adjustment variables, though to a lesser degree. For benchmarks related to family composition and other personal characteristics, variable selection made little difference and proved mildly detrimental for questions of personal finance.

- **The most basic weighting method (raking) performs nearly as well as more elaborate techniques based on matching.** When weighting on both demographic and political variables, methods based on matching resulted in the lowest average bias across the full set of 24 benchmarks – either in combination with raking at smaller sample sizes (n=less than 4,000) or on its own when the sample size was larger. Even so, procedures that only used raking (the least complex method evaluated) performed nearly as well, coming in 0.1 to 0.3 points behind the most effective method, depending on sample size. For benchmarks related to political engagement, the benefits from the more complex approach are somewhat larger than for other topics, doing between 0.5 and 1.2 points better than raking depending on sample size, but nowhere near the magnitude of improvement derived from weighting on political variables

[3] See Dever, Jill A., Ann Rafferty, and Richard Valliant. 2008. "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" Survey Research Methods 2(2), 47-60.

[4] This study examines adjustments that produce a single weight for analyzing all the questions in a survey. It does not consider approaches, such as multilevel regression and post-stratification (MRP), that require a separate model for each question in the survey. The latter may be optimal when there is one outcome of primary interest (e.g., in an election) but can be inefficient for polls exploring a range of topics.

in addition to demographics. If the data necessary to perform matching are readily available and the process can be made routine, then a combination of matching and other methods like raking is likely worthwhile, providing incremental but real improvements.[5] In other situations, such marginal improvements may not be worth the additional statistical labor.

- **Very large sample sizes do not fix the shortcomings of online opt-in samples**. While an online opt-in survey with 8,000 interviews may sound more impressive than one with 2,000, this study finds virtually no difference in accuracy. When adjusting on both demographic and political variables, the most effective procedure at n=8,000 was only 0.2 points better than the most effective procedure at n=2,000. While a large sample size may reduce the variability of estimates (i.e., the modeled margin of error), this is of little help from a "total survey error" perspective. For example, raking on demographic and political variables, the average modeled margin of error across all 24 benchmark variables is ±1.8 percentage points when n=2,000 and ±0.5 points when n=8,000, but the average bias holds steady at 6.3 points. As the sample size increases, estimates become less dispersed and more tightly clustered, but they are often more tightly clustered around the wrong (biased) value.

- **Adjusting on political variables – not just demographics – made key public opinion estimates more Republican.** A prior Pew Research Center study found that online opt-in samples tended to overrepresent Democrats compared with traditional, live telephone random-digit-dial (RDD) samples. In this study, demographic weighting produced almost no change in this distribution or in measures of partisan attitudes such as approval of then-President Barack Obama, views on the Affordable Care Act and 2016 presidential vote. Adding political variables (which include party identification) to the weighting pushes these estimates several points in a Republican direction. For example, support for the Affordable Care Act dropped about 5 percentage points (from 51% to 46%) when the political variables were added to a raking adjustment that initially just used demographics.

The weighting procedures tested in this report represent only a small fraction of the many possible approaches to weighting opt-in survey data. There are a host of different ways to implement matching and propensity weighting, as well as a variety of similar alternatives to raking (collectively known as *calibration* methods). We also did not evaluate methods such as multilevel regression and poststratification, which require a separate statistical model for every outcome variable. Add to this the innumerable combinations of variables that could be used in place of those examined here, and it is clear that there is no shortage of alternative protocols that might have produced different results.

[5] See Dutwin, David and Trent D. Buskirk. 2017. "Apples to Oranges or Gala versus Golden Delicious?: Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples." Public Opinion Quarterly 81(S1), 213-39.

But whatever method one might use, successfully correcting bias in opt-in samples requires having the right adjustment variables. What's more, for at least many of the topics examined here, the "right" adjustment variables include more than the standard set of core demographics. While there can be real, if incremental, benefits from using more sophisticated methods in producing survey estimates, the fact that there was virtually no differentiation between the methods when only demographics were used implies that the use of such methods should not be taken as an indicator of survey accuracy in and of itself. A careful consideration of the factors that differentiate the sample from the population and their association with the survey topic is far more important.

1. How different weighting methods work

Historically, public opinion surveys have relied on the ability to adjust their datasets using a core set of demographics – sex, age, race and ethnicity, educational attainment, and geographic region – to correct any imbalances between the survey sample and the population. These are all variables that are correlated with a broad range of attitudes and behaviors of interest to survey researchers. Additionally, they are well measured on large, high-quality government surveys such as the American Community Survey (ACS), conducted by the U.S. Census Bureau, which means that reliable population benchmarks are readily available.

But are they sufficient for reducing selection bias[6] in online opt-in surveys? Two studies that compared weighted and unweighted estimates from online opt-in samples found that in many instances, demographic weighting only minimally reduced bias, and in some cases actually made bias worse.[7] In a previous Pew Research Center study comparing estimates from nine different online opt-in samples and the probability-based American Trends Panel, the sample that displayed the lowest average bias across 20 benchmarks (Sample I) used a number of variables in its weighting procedure that went beyond basic demographics, and it included factors such as frequency of internet use, voter registration, party identification and ideology.[8] Sample I also employed a more complex statistical process involving three stages: matching followed by a propensity adjustment and finally raking (the techniques are described in detail below).

The present study builds on this prior research and attempts to determine the extent to which the inclusion of different adjustment variables or more sophisticated statistical techniques can improve the quality of estimates from online, opt-in survey samples. For this study, Pew Research Center fielded three large surveys, each with over 10,000 respondents, in June and July of 2016. The surveys each used the same questionnaire, but were fielded with different online, opt-in panel vendors. The vendors were each asked to produce samples with the same demographic distributions (also known as quotas) so that prior to weighting, they would have roughly comparable demographic compositions. The survey included questions on political and social attitudes, news consumption, and religion. It also included a variety of questions drawn from

[6] When survey respondents are self-selected, there is a risk that the resulting sample may differ from the population in ways that bias survey estimates. This is known as selection bias, and it occurs when the kinds of people who choose to participate are systematically different from those who do not on the survey outcomes. Selection bias can occur in both probability-based surveys (in the form of nonresponse) as well as online opt-in surveys.

[7] See Yeager, David S., et al. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." Public Opinion Quarterly 75(4), 709-47; and Gittelman, Steven H., Randall K. Thomas, Paul J. Lavrakas and Victor Lange. 2015. "Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples." Journal of Advertising Research 55(4), 368-79.

[8] In the 2016 Pew Research Center study a standard set of weights based on age, sex, education, race and ethnicity, region, and population density were created for each sample. For samples where vendors provided their own weights, the set of weights that resulted in the lowest average bias was used in the analysis. Only in the case of Sample I did the vendor provide weights resulting in lower bias than the standard weights.

high-quality federal surveys that could be used either for benchmarking purposes or as adjustment variables. (See Appendix A for complete methodological details and Appendix F for the questionnaire.)

This study compares two sets of adjustment variables: core demographics (age, sex, educational attainment, race and Hispanic ethnicity, and census division) and a more expansive set of variables that includes both the core demographic variables and additional variables known to be associated with political attitudes and behaviors. These additional political variables include party identification, ideology, voter registration and identification as an evangelical Christian, and are intended to correct for the higher levels of civic and political engagement and Democratic leaning observed in the Center's previous study.

The analysis compares three primary statistical methods for weighting survey data: raking, matching and propensity weighting. In addition to testing each method individually, we tested four techniques where these methods were applied in different combinations for a total of seven weighting methods:

- Raking
- Matching
- Propensity weighting
- Matching + Propensity weighting
- Matching + Raking
- Propensity weighting+ Raking
- Matching + Propensity weighting + Raking

Because different procedures may be more effective at larger or smaller sample sizes, we simulated survey samples of varying sizes. This was done by taking random subsamples of respondents from each of the three (n=10,000) datasets. The subsample sizes ranged from 2,000 to 8,000 in increments of 500.[9] Each of the weighting methods was applied twice to each simulated survey dataset (subsample): once using only core demographic variables, and once using both demographic and political measures.[10] Despite the use of different vendors, the effects of each weighting protocol were generally consistent across all three samples. Therefore, to simplify reporting, the results presented in this study are averaged across the three samples.

[9] Many surveys feature sample sizes less than 2,000, which raises the question of whether it would be important to simulate smaller sample sizes. For this study, a minimum of 2,000 was chosen so that it would be possible to have 1,500 cases left after performing matching, which involves discarding a portion of the completed interviews.

[10] The process of calculating survey estimates using different weighting procedures was repeated 1,000 times using different randomly selected subsamples. This enabled us to measure the amount of variability introduced by each procedure and distinguish between systematic and random differences in the resulting estimates.

How we combined multiple surveys to create a synthetic model of the population

Often researchers would like to weight data using population targets that come from multiple sources. For instance, the American Community Survey (ACS), conducted by the U.S. Census Bureau, provides high-quality measures of demographics. The Current Population Survey (CPS) Voting and Registration Supplement provides high-quality measures of voter registration. No government surveys measure partisanship, ideology or religious affiliation, but they are measured on surveys such as the General Social Survey (GSS) or Pew Research Center's Religious Landscape Study (RLS).

For some methods, such as raking, this does not present a problem, because they only require summary measures of the population distribution. But other techniques, such as matching or propensity weighting, require a case-level dataset that contains all of the adjustment variables. This is a problem if the variables come from different surveys.

To overcome this challenge, we created a "synthetic" population dataset that took data from the ACS and appended variables from other benchmark surveys (e.g., the CPS and RLS). In this context, "synthetic" means that some of the data came from statistical modeling (imputation) rather than directly from the survey participants' answers.[11]

The first step in this process was to identify the variables that we wanted to append to the ACS, as well as any other questions that the different benchmark surveys had in common. Next, we took the data for these questions from the different benchmark datasets (e.g., the ACS and CPS) and combined them into one large file, with the cases, or interview records, from each survey literally stacked on top of each other. Some of the questions – such as age, sex, race or state – were available on all of the benchmark surveys, but others have large holes with missing data for cases that come from surveys where they were not asked.

[11] The idea for augmenting ACS data with modeled variables from other surveys and measures of its effectiveness can be found in Rivers, Douglas, and Delia Bailey. 2009. "Inference from Matched Samples in the 2008 US National Elections." Presented at the 2009 American Association for Public Opinion Research Annual Conference, Hollywood, Florida; and Ansolabehere, Stephen, and Douglas Rivers. 2013. "Cooperative Survey Research." Annual Review of Political Science 16(1), 307-29.

How to create a synthetic population

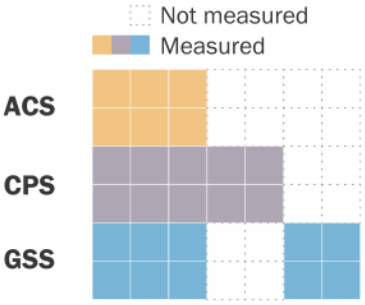
Matching and propensity weighting require case-level data of all adjustment variables. The **American Community Survey** (ACS) provides high-quality measures of demographic data for U.S. adults but lacks other, non-demographic data, including voter registration and ideology, among other variables. These need to be appended from other sources using statistical models. Here is how:

American Community Survey (ACS)

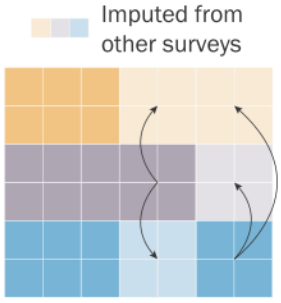
Current Population Survey (CPS)

General Social Survey (GSS)

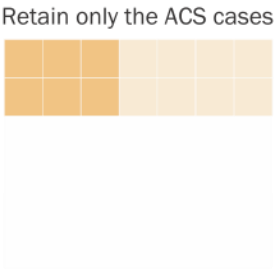
1. First, benchmark datasets containing all common variables, plus those that are to be appended to the ACS are stacked, creating one large dataset.



2. This combined dataset has holes with missing values for items that were measured on some surveys but not others.



3. Next, multiple imputation via chained equations (MICE) is used to fill in the missing values based on any common variables across the different surveys.



4. Finally, all but the cases from the ACS are deleted. This leaves a dataset with the same demographic distribution as the ACS, but augmented with additional, modeled variables that can be used in procedures that need-case level data.

Note: This diagram is intended to provide a simplified overview of the steps taken to create the synthetic population dataset. It depicts only a few of the variables and benchmark surveys used for the study. See Appendix B for a detailed description of the process.
Source: "For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

The next step was to statistically fill the holes of this large but incomplete dataset. For example, all the records from the ACS were missing voter registration, which that survey does not measure. We used a technique called multiple imputation by chained equations (MICE) to fill in such missing

information.[12] MICE fills in likely values based on a statistical model using the common variables. This process is repeated many times, with the model getting more accurate with each iteration. Eventually, all of the cases will have complete data for all of the variables used in the procedure, with the imputed variables following the same multivariate distribution as the surveys where they were actually measured.

The result is a large, case-level dataset that contains all the necessary adjustment variables. For this study, this dataset was then filtered down to only those cases from the ACS. This way, the demographic distribution exactly matches that of the ACS, and the other variables have the values that would be expected given that specific demographic distribution. We refer to this final dataset as the "synthetic population," and it serves as a template or scale model of the total adult population.

This synthetic population dataset was used to perform the matching and the propensity weighting. It was also used as the source for the population distributions used in raking. This approach ensured that all of the weighted survey estimates in the study were based on the same population information. See Appendix B for complete details on the procedure.

Raking

For public opinion surveys, the most prevalent method for weighting is iterative proportional fitting, more commonly referred to as raking. With raking, a researcher chooses a set of variables where the population distribution is known, and the procedure iteratively adjusts the weight for each case until the sample distribution aligns with the population for those variables. For example, a researcher might specify that the sample should be 48% male and 52% female, and 40% with a high school education or less, 31% who have completed some college, and 29% college graduates. The process will adjust the weights so that gender ratio for the weighted survey sample matches the desired population distribution. Next, the weights are adjusted so that the education groups are in the correct proportion. If the adjustment for education pushes the sex distribution out of alignment, then the weights are adjusted again so that men and women are represented in the desired proportion. The process is repeated until the weighted distribution of all of the weighting variables matches their specified targets.

Raking is popular because it is relatively simple to implement, and it only requires knowing the marginal proportions for each variable used in weighting. That is, it is possible to weight on sex, age, education, race and geographic region separately without having to first know the population

[12] See Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?: Multiple Imputation by Chained Equations." International Journal of Methods in Psychiatric Research 20(1), 40–49.

proportion for every combination of characteristics (e.g., the share that are male, 18- to 34-year-old, white college graduates living in the Midwest). Raking is the standard weighting method used by Pew Research Center and many other public pollsters.

In this study, the weighting variables were raked according to their marginal distributions, as well as by two-way cross-classifications for each pair of demographic variables (age, sex, race and ethnicity, education, and region).

Matching

Matching is another technique that has been proposed as a means of adjusting online opt-in samples. It involves starting with a sample of cases (i.e., survey interviews) that is representative of the population and contains all of the variables to be used in the adjustment. This "target" sample serves as a template for what a survey sample would look like if it was randomly selected from the population. In this study, the target samples were selected from our synthetic population dataset, but in practice they could come from other high-quality data sources containing the desired variables. Then, each case in the target sample is paired with the most similar case from the online opt-in sample. When the closest match has been found for all of the cases in the target sample, any unmatched cases from the online opt-in sample are discarded.

If all goes well, the remaining matched cases should be a set that closely resembles the target population. However, there is always a risk that there will be cases in the target sample with no good match in the survey data – instances where the most similar case has very little in common with the target. If there are many such cases, a matched sample may not look much like the target population in the end.

There are a variety of ways both to measure the similarity between individual cases and to perform the matching itself.[13] The procedure employed here used a target sample of 1,500 cases that were randomly selected from the synthetic population dataset. To perform the matching, we temporarily combined the target sample and the online opt-in survey data into a single dataset. Next, we fit a statistical model that uses the adjustment variables (either demographics alone or demographics + political variables) to predict which cases in the combined dataset came from the target sample and which came from the survey data.

The kind of model used was a machine learning procedure called a random forest. Random forests can incorporate a large number of weighting variables and can find complicated relationships

[13] See Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." Statistical Science 25(1), 1-21 for a more technical explanation and review of the many different approaches to matching that have been developed.

between adjustment variables that a researcher may not be aware of in advance. In addition to estimating the probability that each case belongs to either the target sample or the survey, random forests also produce a measure of the similarity between each case and every other case. The random forest similarity measure accounts for how many characteristics two cases have in common (e.g., gender, race and political party) and gives more weight to those variables that best distinguish between cases in the target sample and responses from the survey dataset.[14] We used this similarity measure as the basis for matching.

The final matched sample is selected by sequentially matching each of the 1,500 cases in the target sample to the most similar case in the online opt-in survey dataset. Every subsequent match is restricted to those cases that have not been matched previously. Once the 1,500 best matches have been identified, the remaining survey cases are discarded.

For all of the sample sizes that we simulated for this study (n=2,000 to 8,000), we always matched down to a target sample of 1,500 cases. In simulations that started with a sample of 2,000 cases, 1,500 cases were matched and 500 were discarded. Similarly, for simulations starting with 8,000 cases, 6,500 were discarded. In practice, this would be very wasteful. However, in this case, it enabled us to hold the size of the final matched dataset constant and measure how the effectiveness of matching changes when a larger share of cases is discarded. The larger the starting sample, the more potential matches there are for each case in the target sample – and, hopefully, the lower the chances of poor-quality matches.

Propensity weighting

A key concept in probability-based sampling is that if survey respondents have different probabilities of selection, weighting each case by the *inverse* of its probability of selection removes any bias that might result from having different kinds of people represented in the wrong proportion. The same principle applies to online opt-in samples. The only difference is that for probability-based surveys, the selection probabilities are known from the sample design, while for opt-in surveys they are unknown and can only be estimated.

For this study, these probabilities were estimated by combining the online opt-in sample with the entire synthetic population dataset and fitting a statistical model to estimate the probability that a case comes from the synthetic population dataset or the online opt-in sample. As with matching, random forests were used to calculate these probabilities, but this can also be done with other

[14] See Appendix C for a more detailed explanation of random forests and the matching algorithm used in this report, as well as Zhao, Peng, Xiaogang Su, Tingting Ge and Juanjuan Fan. 2016. "Propensity Score and Proximity Matching Using Random Forest." Contemporary Clinical Trials 47, 85-92.

kinds of models, such as logistic regression.[15] Each online opt-in case was given a weight equal to the estimated probability that it came from the synthetic population divided by the estimated probability that it came from the online opt-in sample. Cases with a low probability of being from the online opt-in sample were underrepresented relative to their share of the population and received large weights. Cases with a high probability were overrepresented and received lower weights.

As with matching, the use of a random forest model should mean that interactions or complex relationships in the data are automatically detected and accounted for in the weights. However, unlike matching, none of the cases are thrown away. A potential disadvantage of the propensity approach is the possibility of highly variable weights, which can lead to greater variability for estimates (e.g., larger margins of error).

Combinations of adjustments

Some studies have found that a first stage of adjustment using matching or propensity weighting followed by a second stage of adjustment using raking can be more effective in reducing bias than any single method applied on its own.[16] Neither matching nor propensity weighting will force the sample to exactly match the population on all dimensions, but the random forest models used to create these weights may pick up on relationships between the adjustment variables that raking would miss. Following up with raking may keep those relationships in place while bringing the sample fully into alignment with the population margins.

These procedures work by using the output from earlier stages as the input to later stages. For example, for matching followed by raking (M+R), raking is applied only the 1,500 matched cases. For matching followed by propensity weighting (M+P), the 1,500 matched cases are combined with the 1,500 records in the target sample. The propensity model is then fit to these 3,000 cases, and the resulting scores are used to create weights for the matched cases. When this is followed by a third stage of raking (M+P+R), the propensity weights are trimmed and then used as the starting point in the raking process. When first-stage propensity weights are followed by raking (P+R), the process is the same, with the propensity weights being trimmed and then fed into the raking procedure.

[15] See Buskirk, Trent D., and Stanislav Kolenikov. 2015. "Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification." Survey Methods: Insights from the Field (SMIF).
[16] See Dutwin, David and Trent D. Buskirk. 2017. "Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples." Public Opinion Quarterly 81(S1), 213-239.

2. Reducing bias on benchmarks

To understand the relative merits of alternative adjustment procedures, each was assessed on its effectiveness at reducing bias for 24 different benchmarks drawn from high-quality, "gold-standard" surveys. These benchmarks cover a range of topics including civic and political engagement (both difficult topics for surveys in general), technology use, personal finances, household composition and other personal characteristics. See Appendix D for a complete list. While these benchmarks all come from high-quality surveys, it is important to note that these measures are themselves estimates and are subject to error. As a result, the estimates of bias described here should be thought of as approximations.

For each simulated survey dataset with sample sizes ranging from 2,000 to 8,000, the seven statistical techniques were applied twice, once using only demographic variables, and once using both demographic and political variables. This produced a total of 14 different sets of weights for each dataset. Next, estimates were calculated for each substantive category[17] of the 24 benchmark questions using each set of weights as well as unweighted.

The estimated bias for each category is the difference between the survey estimate and the benchmark value.[18] To summarize the level of bias for all of the categories of a particular

Topics and corresponding benchmarks

Topic	Benchmark
Civic engagement	How often talks with neighbors
	Trusts neighbors
	Participated in a school group, neighborhood, or community association
	Volunteered in past year
Family	Marital status
	Presence of children in household
	Household size
Financial	Employment status
	Home ownership
	Family income
	Household member received food stamps
	Health insurance
Personal	Lived in house or apartment one year ago
	Active duty military service
	U.S. citizenship
	Gun ownership
	Smoking
	Food allergies
Political engagement	Voted in 2012
	Voted in 2014
	Contacted or visited a public official in past year
Technology	Tablet or e-reader use
	Texting or instant messaging
	Social networking

Note: See Appendix D for the source of each benchmark, the question text, the response categories, the benchmark estimate, and additional notes.
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

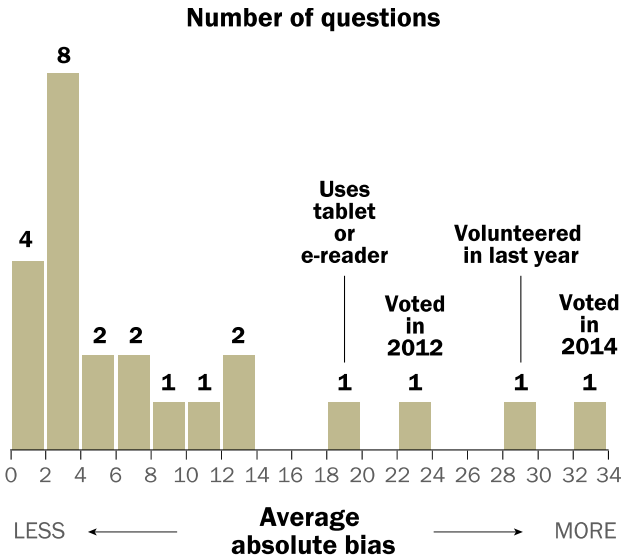
[17] When present, item nonresponse categories such as "Refused" or "Don't know" were included in the base when calculating percentages but were otherwise ignored in the analysis.
[18] The full analysis was repeated 1,000 times using different randomly selected subsamples. Point estimates are calculated as the average value over 1,000 replications. Bias is estimated as the average difference between the survey estimate and the benchmark value over 1,000 replications.

benchmark variable, we calculated the average of the absolute values of the estimated biases for each of the variable's categories. To summarize the overall level of bias across multiple questions (e.g., all 24 benchmarks), the average of the question-level averages was used.

Prior to any weighting, the estimated average absolute bias for the 24 benchmark variables was 8.4 percentage points. Many of the estimated biases are relatively small. Half of the variables have average biases under 4 points, four of which are under 2 points (family income, home ownership, marital status and health insurance coverage). At the other end of the scale, four variables show extremely large biases. These are voting in the 2014 midterm election, (32 percentage points), having volunteered in the past 12 months (29 points), voting in the 2012 presidential election (23 points) and tablet ownership (20 points).

Tablet use, volunteering and voting exhibit the largest biases

Number of benchmark questions at different levels of average absolute bias (percentage points)



Note: Measures of bias depicted in this figure are unweighted.
Source: Pew Research Center analysis of three online opt-in surveys. "For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

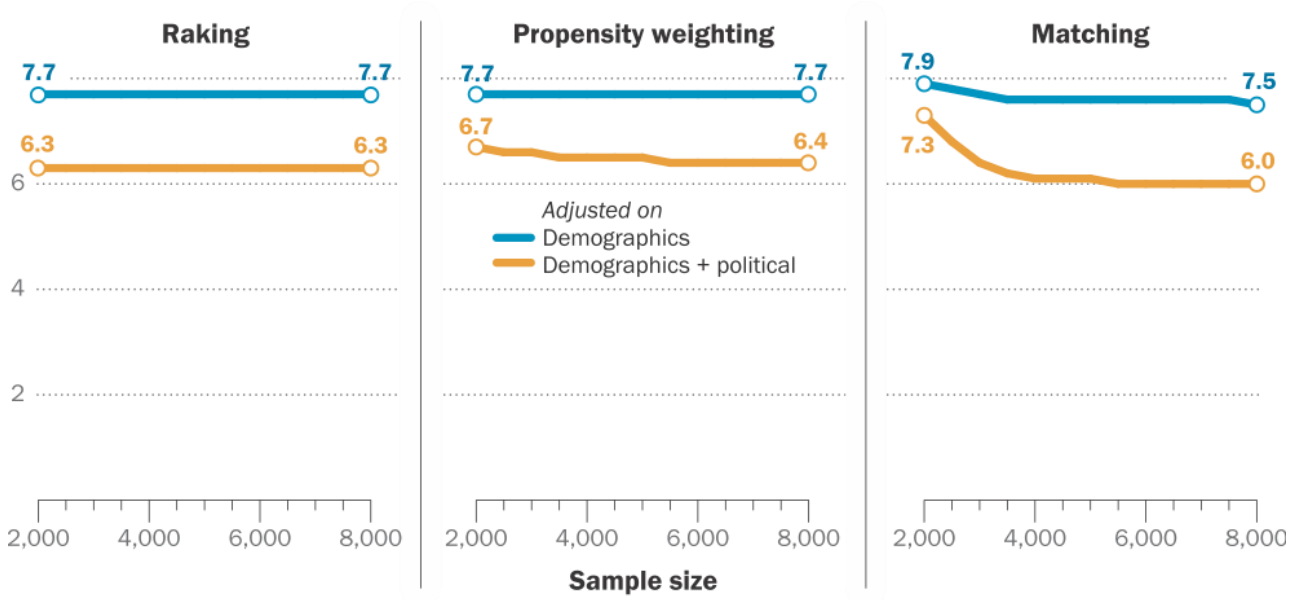
Choice of variables for weighting is more consequential than the statistical method

More than any other factor, the choice of adjustment variables has the largest impact on the accuracy of estimates. Adjusting on both the demographic and political variables resulted in lower average bias than adjusting on demographics alone. While the largest improvements were for measures of political engagement (such as voting), benchmarks related to civic engagement and technology use also saw sizable reductions in bias. The differences between survey topics are examined in detail in the section "Results by question topic reflect correlations with the adjustment variables."

This was true for all three primary statistical methods as well as the four combination methods, and it was true at every sample size. On average, adjusting on demographics alone reduced estimated bias by just under 1 percentage point, from 8.4 points before weighting to 7.6 after. This effect was relatively consistent regardless of the statistical method or sample size. By contrast, weighting on both demographics and the political variables reduces bias an additional 1.4

Using both demographic and political variables resulted in lower bias across all three primary methods

Average absolute differences between population benchmarks and weighted sample estimates (percentage points)



Source: Pew Research Center analysis of three online opt-in surveys.
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

percentage points on average, although the degree of improvement was more sensitive to the statistical method and sample size. Under the best-case scenario, the more comprehensive set of adjustment variables reduced the average estimated bias to a low of 6 percentage points.

Matching on its own can improve upon raking, but only modestly and with large samples

The study examined how the performance of each adjustment method is affected by sample size. For raking, the reduction in bias was effectively the same at every sample size. The average estimated bias with n=8,000 interviews is identical to that with n=2,000 interviews (7.7 percentage points when adjusting on demographics and 6.3 for demographic + political variables).

Matching, on the other hand, becomes more effective with larger starting sample sizes because there are more match candidates for each case in the target sample. When adjusting on demographic variables, matching did show a small improvement as the sample size increased, going from an average estimated bias of 7.9 percentage points at a starting sample size of n=2,000 to 7.5 points at n=8,000. When political variables were included in the adjustment, the benefits of

a larger starting sample are more substantial, going from a high of 7.3 points at n=2,000 to a low of 6 points at n=8,000. Even so, matching reached a point of diminishing returns around n=4,000 and levels off completely at n=5,500 and greater. This suggests that that there may not be much benefit from further increasing the size of the starting sample when the target size is 1,500.

Most notably, matching by itself performed quite poorly relative to raking at smaller sample sizes. When n=2,000, raking's average estimated bias was a full point lower for raking. Matching did not overtake raking until the starting sample size reached 3,500. At best, matching improved upon raking by a relatively modest 0.3 points, and then only at sample sizes of 5,500 or larger. None of the opt-in panel vendors that regularly employ matching use this approach on its own; rather, they follow matching with additional stages of adjustment or statistical modeling.

Unlike matching, propensity weighting was never more effective than raking. When only demographics were used, the estimated bias was equal to raking at a constant at 7.7 percentage points. With both demographic and political variables employed, propensity-weighting bias ranged from 6.7 points when n=2,000 to 6.4 points at n=8,000. This improvement likely occurs because the random forest algorithm used to estimate the propensities can fit more complex and powerful models given more data and more variables.

Raking in addition to matching or propensity weighting can be better than raking alone

When multiple techniques were used together in sequence, the result was slightly more bias correction than any of the methods on their own. At smaller starting sample sizes (e.g., n=less than 4,000), matching performed quite poorly relative to raking. But if both matching and raking were performed, the result was slightly lower bias than with raking alone. For example, when a starting sample of n=2,000 was matched on both demographic and political variables, the average estimated bias was 7.3 points, but when the matching was followed by raking, the average bias dropped to 6.2 points, putting it just ahead of raking by 0.1 point on average.

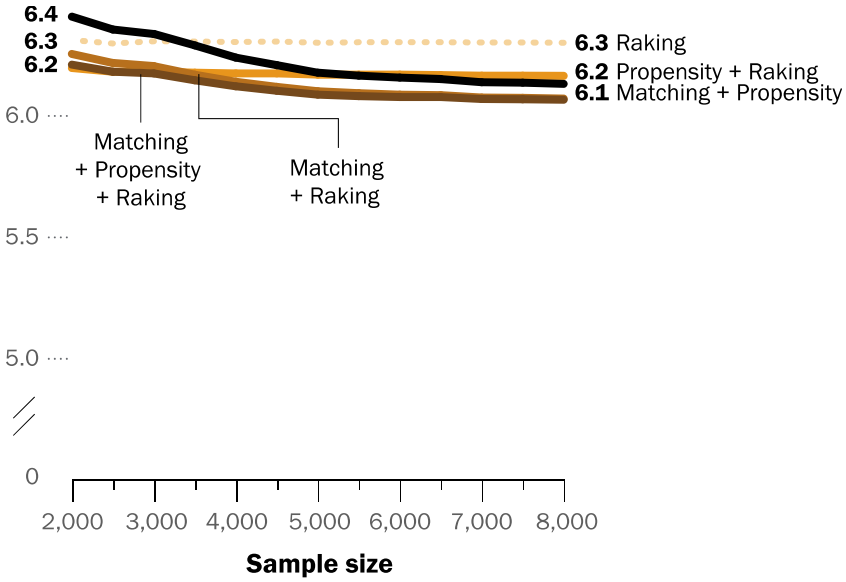
When matching was followed by propensity weighting, there was some improvement in accuracy, but not as much. A third stage of raking applied after propensity weighting produced the same results as just matching plus raking, suggesting that any added benefit from an intermediate propensity weighting step is minimal.

A similar pattern emerged when propensity weighting was followed by raking. On its own, propensity weighting always performed worse than raking, but when the two were used in combination with both demographic and political variables, the result was a small but consistent

improvement of 0.1 points compared with raking alone. While there are few scenarios where matching or propensity weighting would be preferable to raking when used in isolation, they can add value when combined with raking. That being said, the benefits are very small, on the order of 0.1 percentage points, and may not be worth the extra effort.

Combining several methods performed slightly better than raking on its own

Average absolute differences between population benchmarks and weighted sample estimates (percentage points)



Note: Figures are based on adjustments performed using both demographic and political variables.
Source: Pew Research Center analysis of three online opt-in surveys. "For Weighting Online Opt-In Samples, What Matters Most?"

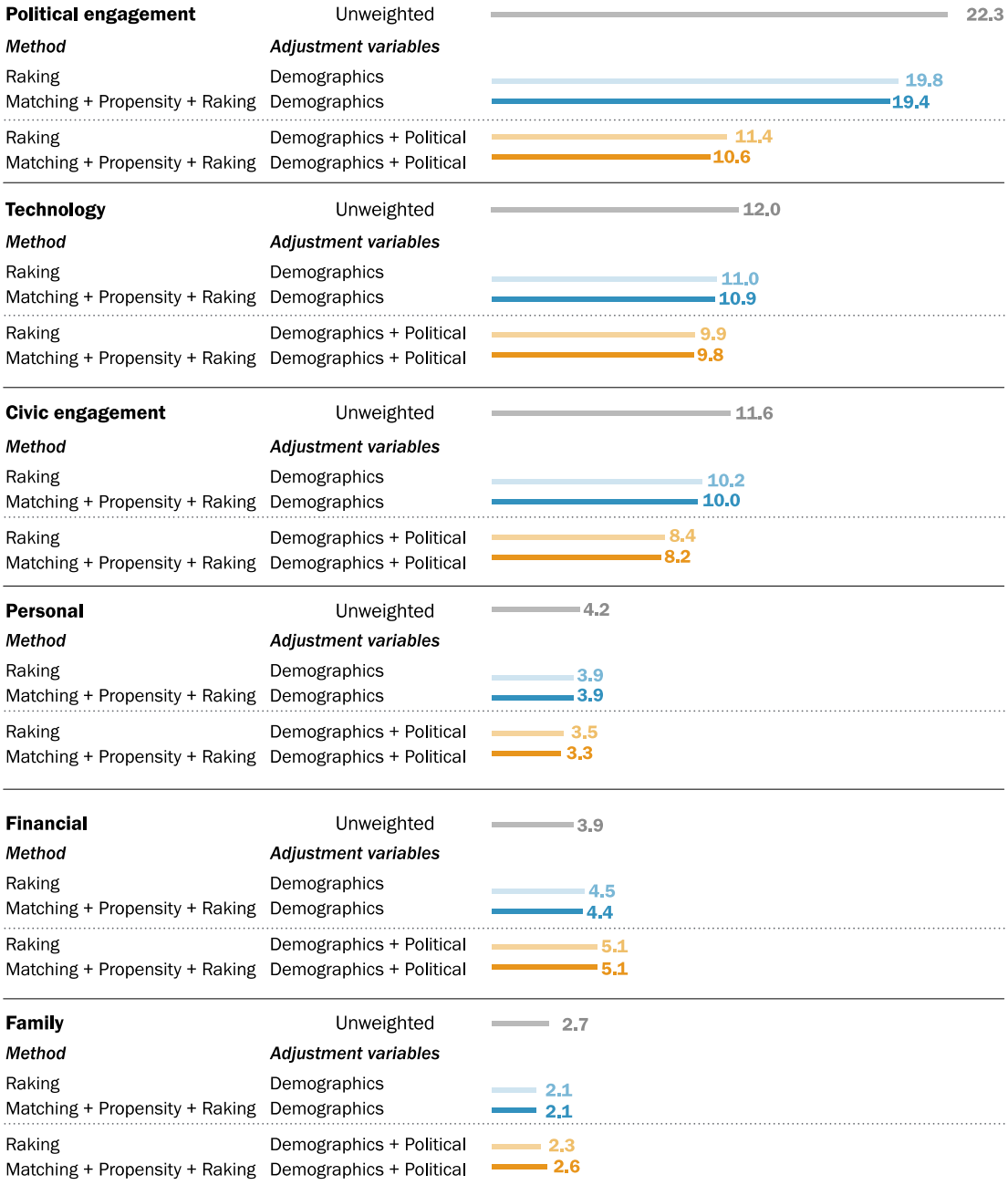
PEW RESEARCH CENTER

Perhaps the most interesting finding was how little benefit came from having a large sample size. The most effective adjustment protocol reduced the average bias to 6 percentage points with a sample size of at least n=5,500, only 0.2 points better than can be achieved with n=2,000. Why does the average estimated bias plateau at about 6 percentage points? Why doesn't the bias level keep declining toward zero as the sample size goes to n=8,000? The survey literature suggests this is because the more comprehensive set of adjustment variables (i.e., the nine demographic + political variables) still does not fully capture the ways in which the online opt-in respondents differ from the population of U.S. adults.[19] In other words, there are other characteristics, which have not been identified, on which the online opt-in sample differs from the population, and those differences result in bias, even after elaborate weighting adjustments are applied. Increasing the sample size to 8,000 does not solve this problem, because the additional interviews are just "more of the same" kinds of adults with respect to the adjustment variables and survey outcomes.

[19] See Mercer, Andrew W., Frauke Kreuter, Scott Keeter, and Elizabeth A. Stuart. 2017. "Theory and Practice in Nonprobability Surveys: Parallels Between Causal Inference and Survey Inference." Public Opinion Quarterly 81(S1), 250-71.

Political engagement estimates see the largest improvement from weighting on both demographic and political variables

Average absolute differences between population benchmarks and sample estimates by survey topic (percentage pts.)



Note: Figures are based on a simulated sample size of 3,500.
Source: Pew Research Center analysis of three online opt-in surveys. "For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Choice of adjustment variables has a much larger impact when related to the survey topic

In terms of improving the accuracy of estimates, the results for individual survey topics (e.g., personal finances, technology, household characteristics) were similar to what was observed in the aggregate. Specifically, the study finds that the choice of adjustment variables mattered much more than the choice of statistical method. That said, the effect varied considerably from topic to topic.

The example comparing the least complex method, raking, to the more elaborate approach of matching followed by propensity adjustment and raking ("M+P+R") across topics is illustrative of the general pattern. For the political engagement topic, M+P+R resulted in slightly lower bias than raking with both sets of adjustment variables, but the two methods were largely indistinguishable for the remaining topics. Meanwhile, the difference between adjusting on demographics alone and including additional political variables can be substantial. The difference was most dramatic for political engagement, which had an average bias of 22.3 percentage points unweighted – higher than any other topic. M+P+R with demographic variables reduced this by 2.9 points, but the inclusion of political variables reduced the average bias by an additional 8.8 points.

For political engagement benchmarks, the unweighted estimates substantially overrepresented adults who voted in 2014 and 2012 by 32 and 23 percentage points respectively. While M+P+R with demographic variables reduced these biases somewhat (by 3 and 4 points for the respective voting years), the inclusion of political variables in adjustment reduced the bias in the 2012 and 2014 votes by an additional 11 and 12 points respectively. This is likely due to the inclusion of voter registration as one of the political adjustment variables. Prior to weighting, registered voters were overrepresented by 19 percentage points, and it is natural that weighting registered voters down to their population proportion would also bring down the share who report having voted. The reduction in estimated bias on the share who reported contacting or visiting a public official in the past year makes intuitive sense as well, since it is plausible that those individuals are also more likely to be registered to vote.

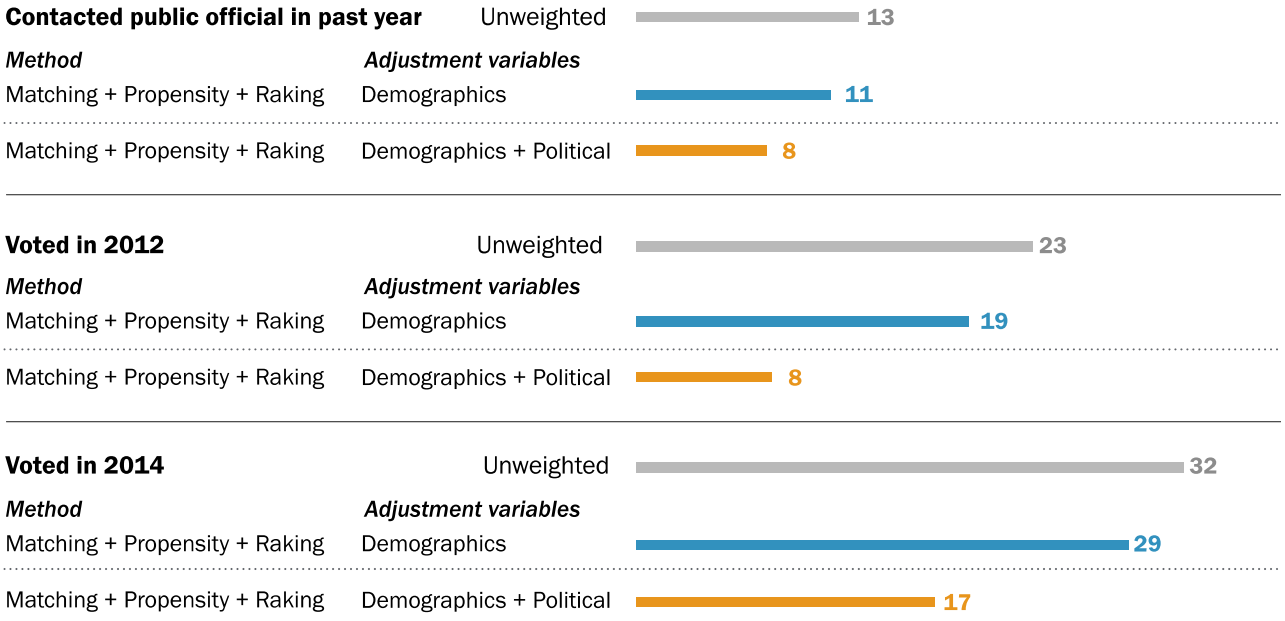
However, even though the addition of political variables corrected a great deal of bias on these measures, large biases remained, with voting in 2012 overestimated by 8 points and voting in 2014 overestimated by 17 points. It is very possible that at least some of this remaining bias reflects individuals claiming to have voted when they did not, either because they forgot or because voting is socially desirable. Either way, the use of political variables in adjustment is not a silver bullet.

The effect of adjustment on questions about personal finances merits particular attention. On these questions, weighting caused the average estimated bias to *increase* rather than decrease, and the use of the expanded political variables made the increase even larger. Prior to any adjustment, the samples tend toward lower levels of economic well-being than the general public. For example, individuals with annual household incomes of $100,000 or more were underrepresented by about 8 percentage points, while those with incomes of under $20,000 were overrepresented by about 4 points. The share of respondents employed full-time was about 6 points lower than the population benchmark, while the percentage unemployed, laid off or looking for work was almost 5 points higher than among the population. The percentage who report that a member of their household has received food stamps in the past year was 13 points higher than the benchmark.

At the same time, respondents tended to have *higher* levels of education than the general public. The unweighted share with postgraduate degrees was 6 percentage points higher than the population value, and the percentage with less than a high school education 8 points lower.

Improvement in political engagement estimates was driven by large corrections on measures of voting

Average differences between population benchmarks and sample estimates (percentage points)



Note: Figures are based on a simulated sample size of 3,500.
Source: Pew Research Center analysis of three online opt-in surveys. "For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Adjusting on core demographic variables corrected this educational imbalance and reduced the average education level of the survey samples. But in doing so, the average level of economic well-being was reduced even further, and biases on the financial measures were magnified rather than reduced. Because financial well-being and voter registration are also positively correlated, the inclusion of the expanded political variables produces even larger biases for these variables. This pattern suggests that weighting procedures could benefit from the inclusion of one or more additional variables that capture respondents' economic situations more directly than education.

For benchmarks pertaining to civic engagement and technology, the reduction in bias from the inclusion of political variables was just over twice that of demographics alone, although in both cases the reductions were smaller than for political engagement. On the other hand, bias reduction for the personal and family topics was minimal for both sets of variables.

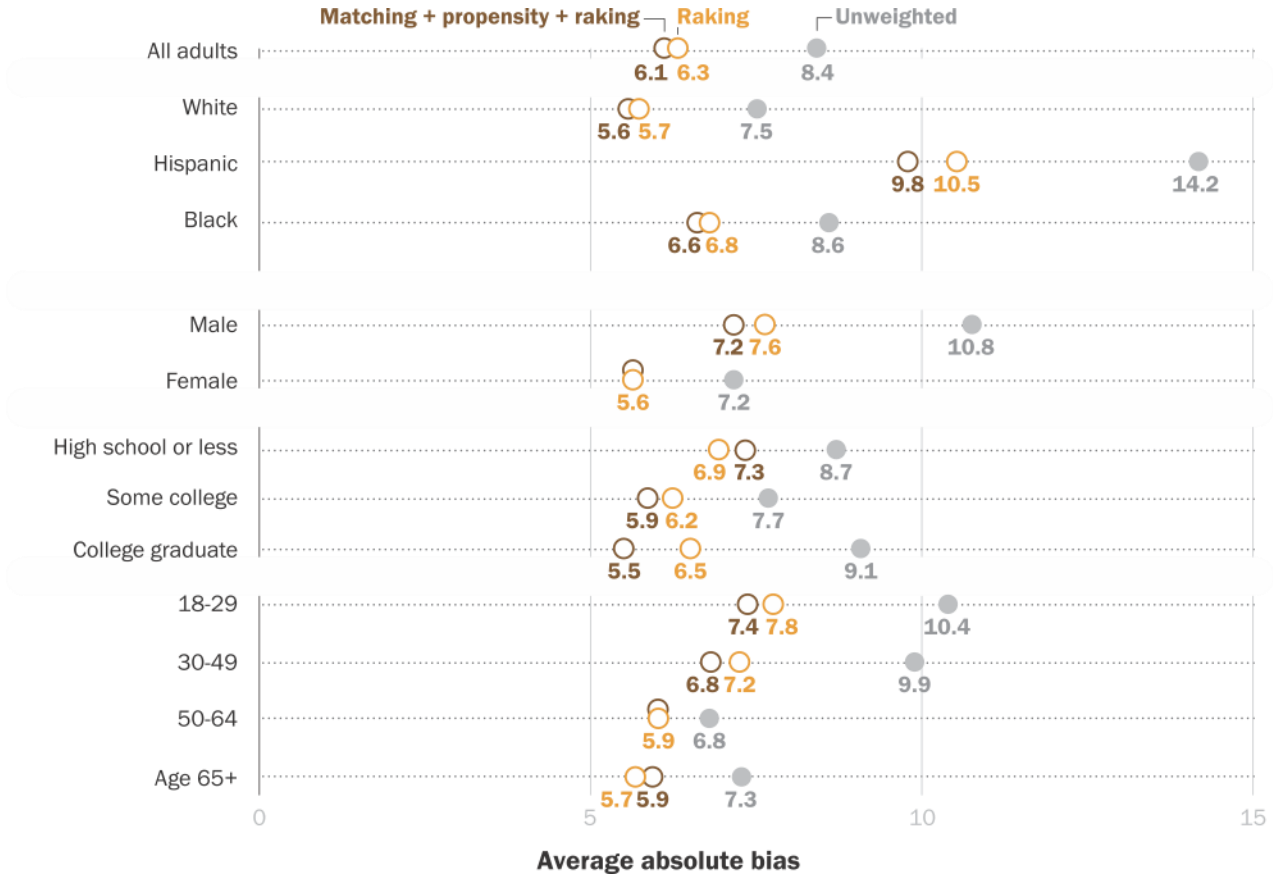
For some subgroups, more elaborate adjustments outperform raking

While there may be little to be gained from very large sample sizes or more complex statistical methods as far as general population estimates are concerned, there could be more pronounced differences between adjustment methods or more of an impact from increasing sample size for survey estimates based on population subgroups. In fact, an appealing feature of the machine-learning models used in matching and propensity weighting is the possibility that they will detect imbalances within subgroups that a researcher might not think to account for with raking.

For most subgroups, raking performed nearly as well as more elaborate approaches. However, there were a few subgroups that saw somewhat larger improvements in accuracy with more complex approaches. To minimize the number of moving parts in this particular analysis, these results are all based on a sample size of n=3,500 and on adjustments using both the demographic and political variables. Estimates based on college graduates had an average estimated bias of 6.5 percentage points with raking versus 5.5 points with a combination of matching, propensity weighting and raking. Similarly, average estimated bias on Hispanic estimates was 10.5 percentage points with raking versus 9.8 with the combination method. Similar, though smaller, differences were found for estimates based on adults ages 18-29, those ages 30-49, and men. Conversely, estimates for those with no more than a high school education were somewhat more accurate with raking. Estimates for other major demographic subgroups did not appear to be affected by the choice of statistical method.

Estimates for college graduates and Hispanics see a larger benefit from more complex methods

Average absolute differences between population benchmarks and sample estimates among … (percentage points)



Note: Figures are based on adjustments performed using both demographic and political variables and a simulated sample size of 3,500.
Source: Pew Research Center analysis of three online opt-in surveys. All respondents who identified themselves as having Hispanic ethnicity were classified as Hispanic, regardless of what race they identified as.
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

The pattern for Hispanics is particularly noteworthy. Estimates for this group had the largest average bias, both before weighting and after. The fact that M+P+R performed better than raking suggests that there are imbalances in the Hispanic composition that are not sufficiently captured by the raking specification. While this was the case for other groups as well (e.g., college graduates), Hispanics also saw much larger benefits from a larger starting sample size than other subgroups. At n=2,000, the average bias for Hispanic estimates was 10.2 percentage points. This steadily declined to 9 points at n=8,000 without leveling off, for a total change of 1.2 points. In comparison, the next-largest shifts were observed for college graduates, men, and adults younger

than 30, at 0.4 points. This implies that even with 8,000 cases to choose from, the quality of Hispanic matches was poor and problematic in ways that subsequent propensity weighting and raking steps were unable to overcome. While all subgroups exhibited biases, the representation of Hispanics is particularly challenging and will require additional efforts that go well beyond those tested in this study.

For partisan measures, adding political variables to weighting adjustment can make online opt-in estimates more Republican

While benchmark comparisons provide an important measure of data quality, public opinion researchers are usually interested in studying attitudes and behaviors that lack the same kind of ground truth that can be used to gauge their accuracy. When gold-standard benchmarks are not available, one way to assess online opt-in polls is to look for alignment with probability-based polls conducted at roughly the same point in time. Although these polls are not without flaws of their own, well-designed and executed probability-based methods tend to be more accurate.[20]

In this study, there were several measures that could be compared to contemporaneous public polling: Barack Obama's presidential approval, attitudes about the Affordable Care Act, and presidential vote preference in the 2016 election. These kinds of partisan measures are particularly relevant given that a previous Pew Research Center study found that online opt-in samples ranged from 3 to 8 percentage points more Democratic than comparable RDD telephone surveys.

The surveys used in this study showed a similar pattern. The synthetic population dataset had a distribution of 30% Democrat, 22% Republican and 48% independent or some other party, very close to the distribution found on the GSS and Pew Research Center surveys used in its creation. In comparison, with demographics-only raking, the opt-in samples used in this study were on average 4 points more Republican and 8 points more Democratic than the synthetic population dataset – more partisan in general, but disproportionately favoring Democrats. This is almost identical to the partisan distribution without any weighting at all.

[20] While tight correspondence with results from probability-based polls is suggestive of accuracy, this type of analysis does not support estimation of bias in the way that benchmarking does. Even fairly rigorous probability-based opinion polls are apt to contain too much error (e.g., from sampling, nonresponse or measurement) to be treated as precise population benchmarks.

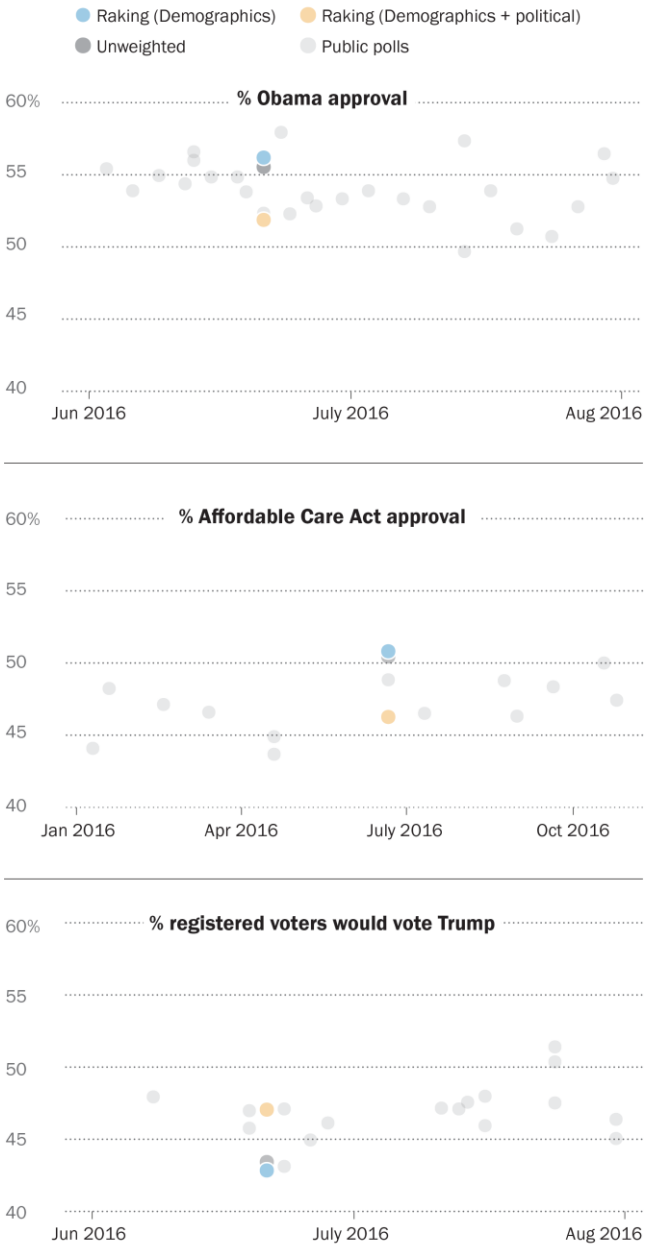
Using the political variables (which include party identification) in addition to demographics brings partisanship in line with the synthetic frame, reducing the share of Democrats more than the share of Republicans, and substantially increasing the share of independents.

This has a commensurate effect on public opinion measures that are associated with partisanship, moving them several points in the Republican direction. For example, when raking on demographics only, Obama's approval rating was 56%, while adding political variables reduced this to 52%. Similarly, support for the Affordable Care Act dropped about 5 percentage points (from 51% to 46%) when the political variables were added to the raking adjustment. Support for Donald Trump among registered voters increased 4 points (from 43% to 47%) when the political variables were added.[21]

This raises the important question of whether these shifts in a Republican direction represent an improvement in data quality. The demographically weighted estimates do appear to be more Democratic than the probability-based polling with respect to each of these three measures, although in each case, they are not so different as to be entirely implausible. Although it is not possible to say definitively that the estimates that adjust on both

Adjusting on both demographic and political variables made attitudinal estimates more Republican

Online opt-in estimates compared with estimates from probability-based public polls



Note: For consistency, measures of support for Trump are limited to surveys that asked undecided voters if they leaned toward one of the candidates. Estimates from online opt-in samples are based on a simulated sample size of 3,500.
Source: Public polling conducted by Pew Research Center, CBS News, CNN, Gallup, ABC/Washington Post, Bloomberg/Selzer, Kaiser Family Foundation, CBS/New York Times, Fox News, University of Delaware/PSRAI, Monmouth University, NBC/Wall Street Journal, and Associated Press/GfK; Pew Research Center analysis of three online opt-in samples.
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

[21] On these outcomes, there was very little difference between adjusting on demographics versus not weighting the data at all. There were also no major differences between different statistical methods. For example, Obama's approval was 52% using either raking or the combination of matching, propensity adjustment and raking (when both demographic and political variables were adjusted for).

demographic and political variables are more accurate, they do appear to be more in line with the trends observed in the other surveys.

While there appears to be a partisan tilt in many online opt-in surveys that should be addressed, particularly research focused on political topics, a great deal of caution is warranted. The partisan distribution of the American public changes over time, and the use of out-of-date weighting parameters could hide real changes in public opinion.

Less partisan, more ideological measures showed smaller changes when political variables were added to weighting

This study included a number of other attitudinal measures for which comparisons to other public polling was not possible. Nevertheless, it is still helpful to see the extent to which they are sensitive to decisions about weighting procedures. Many of these measures capture ideological but not necessarily partisan differences. Adjusting on both demographic and political variables tended to shift these measures in a more conservative direction, though the effect was both less pronounced and less consistent than for overtly partisan measures. For simplicity, the discussion is limited to estimates produced via raking, though as with the other attitudinal measures, there were no discernible differences from estimates employing more complex methods.

Using political variables in addition to demographics brought the percentage who said there is a lot of discrimination against blacks from 58% to 55%; against gays and lesbians, from 60% to 57%; and against Hispanics, from 52% to 49%. Support for marijuana legalization decreased from 61% to 58%.

The adjustment's effects on other attitudinal estimates were more muted. The percentage who agreed with the statement "Immigrants today strengthen our country because of their hard work and talents" was essentially unchanged, going from 51% to 50%. The percentage who agreed that "Government should do more to solve problems" was 56% for demographics and 55% for demographics + political variables. The percentage who agreed with the statement "The economic system in this country unfairly favors powerful interests" stayed at 72%, regardless of whether only demographics or both demographics and political variables were used for adjustment. The share saying race relations in the United States were "getting better" went from 21% to 19%.

The study also contained questions about respondents' engagement with public affairs and with the news. Adjusting on both demographic and political variables made these online opt-in estimates less engaged, with the percentage who would say they follow what's going on in government and public affairs "most of the time" decreasing from 38% to 34% and the percentage

who would say they follow the news “all or most of the time” falling from 48% to 44%. These somewhat larger shifts are in line with the reductions in political and civic engagement that were observed on benchmarks.

3. Variability of survey estimates

While previous sections of this report have focused on the kinds of systematic biases that may be the largest worry when it comes to public opinion polls, the variance (or precision) of estimates is important as well. Pollsters most commonly talk about precision in terms of the "margin of error" (MOE), which describes how much survey estimates are expected to bounce around if one were to repeat the survey many times identically. For probability-based surveys, the margin of error is usually based on the inherent mathematical properties of random samples. For opt-in samples, this is not possible. Instead, the MOE must be based on modeling assumptions about what other hypothetical samples would look like if the same sampling process were repeated many times. Although the interpretation is largely the same as for probability-based samples, we call it a "modeled" margin of error in order to explicitly acknowledge the reliance on these assumptions.[22]

This kind of error is *in addition to* any systematic biases caused by noncoverage, nonresponse or self-selection. For instance, an estimate with a MOE of ±3 percentage points and no bias would usually fall within 3 points of the truth. If the bias were +10 points, the same margin of error would mean that the estimates would usually fall 7 to 13 points higher than the truth – spread out in the same way but centered on the wrong value.

While sample size is usually considered the largest factor in determining the MOE, survey precision is also affected by weighting. Including more variables in adjustment usually leads to a larger MOE, as does throwing away observations when performing matching.

To see how different procedures influence variability, we calculated the modeled MOE for each of the 81 estimates from all 24 benchmark variables and took the average.[23] Unweighted, the average margin of error on the benchmarks was ±1.3 percentage points for a sample size of n=2,000. As the sample size increased, the average MOE shrank to a low of ±0.4 points at n=8,000.

The modeled margin of error increases only slightly with the addition of political variables

One clear finding is that the use of the political variables in addition to basic demographics has a minimal effect on the margin of error. For all 14 methods and across every sample size, adding

[22] In this case, we are assuming that the sampling process for these surveys is similar to a simple random sample of the vendors' panels, and that repeated surveys would produce samples with similar characteristics prior to any weighting. This assumption may underestimate the amount of variability that would occur in practice if a survey was repeated using the same panel, but it provides a reasonable baseline against which to measure the relative effects of different adjustment procedures on variability.

[23] Here, "margin of error" refers to the half-width of a 95% confidence interval for a percentage. The upper and lower bounds of the confidence intervals were calculated using the 97.5th and 2.5th percentile values for each estimate over 1,000 replications.

political variables to the adjustment procedure never increased the average MOE by more than 0.2 percentage points. In most cases, the difference was even smaller, and in some cases the average

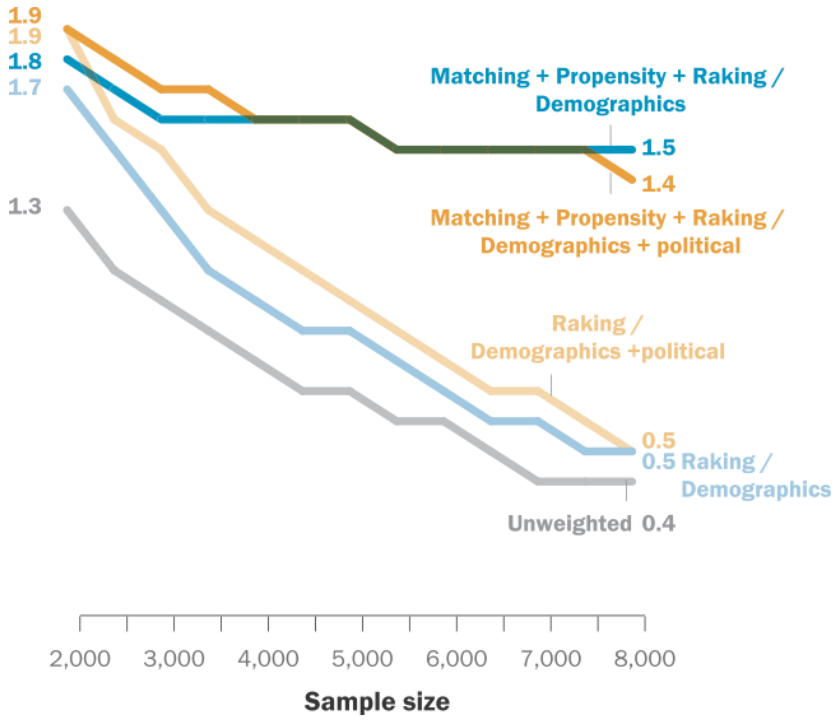
MOE was actually smaller with the political variables than without.[24] Given this consistent pattern, the remainder of this section will focus only on procedures that adjust on both demographic and political variables.

The loss of precision from matching starts out small but increases quickly with sample size

At smaller sample sizes, the choice of statistical method also has a relatively small effect on the precision of estimates. When n=2,000, the four most effective methods for reducing bias (raking plus the combination methods that use raking as the final stage: P+R, M+R, and M+P+R) all have an average margin of error of ±1.9 percentage points. The other combination method, matching followed by propensity weighting (M+P), is very close, at ±1.8 points. Matching and propensity weighting on their own show somewhat lower MOEs at ±1.6 and ±1.5 percentage points respectively – a modest improvement

For the margin of error, method matters less with smaller samples

Average modeled margin of error (± percentage points)



Note: The modeled margin of error for each estimate is calculated as one half the width of a 95% confidence interval. This was calculated for all 81 substantive categories from 24 benchmark items and then averaged.
Source: Pew Research Center analysis of three online opt-in surveys.
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

[24] Although this finding runs counter to the normal expectation that adjusting on a larger number of variables leads to larger margins of error, weighting can actually reduce variability for some estimates if the weights are strongly correlated with the outcomes. This will naturally be the case for estimates that see a large change in the bias (either positive or negative) when the variables are added. For additional details see Little, Roderick J., and Sonya L. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" Survey Methodology 31 (2), 4-11.

but unlikely enough to make up for the fact that these methods performed comparatively poorly with respect to bias.

The fact that two methods that retain all of the interviews (raking and P+R) can have the same average MOE as two for which a quarter of the interviews are discarded (M+R and M+P+R) is perhaps surprising, though it serves to highlight the different trade-offs that are involved with each approach. For the former, estimates use the full sample size, but bias reduction is achieved through more variable and extreme weights, which tends to increase the variance of survey estimates. For the latter, estimates use only the matched 1,500 cases, but the weights generated by the subsequent propensity weighting and raking steps are less extreme.

However, as the starting sample size increases, so does the share of interviews that are discarded in the matching process, and the resulting penalty quickly becomes large relative to methods that retain all of the interviews. In this study, by the time the sample size reached 8,000, the methods that retained all interviews (raking and P+R) both had an average MOE of ±0.5. In contrast, the MOE for the two matching methods (M+R and M+P+R) only fell to ±1.4 at that size. Notably, the use of propensity weighting as either the first or second step appeared to have little to no effect on the average margin of error when followed by raking.

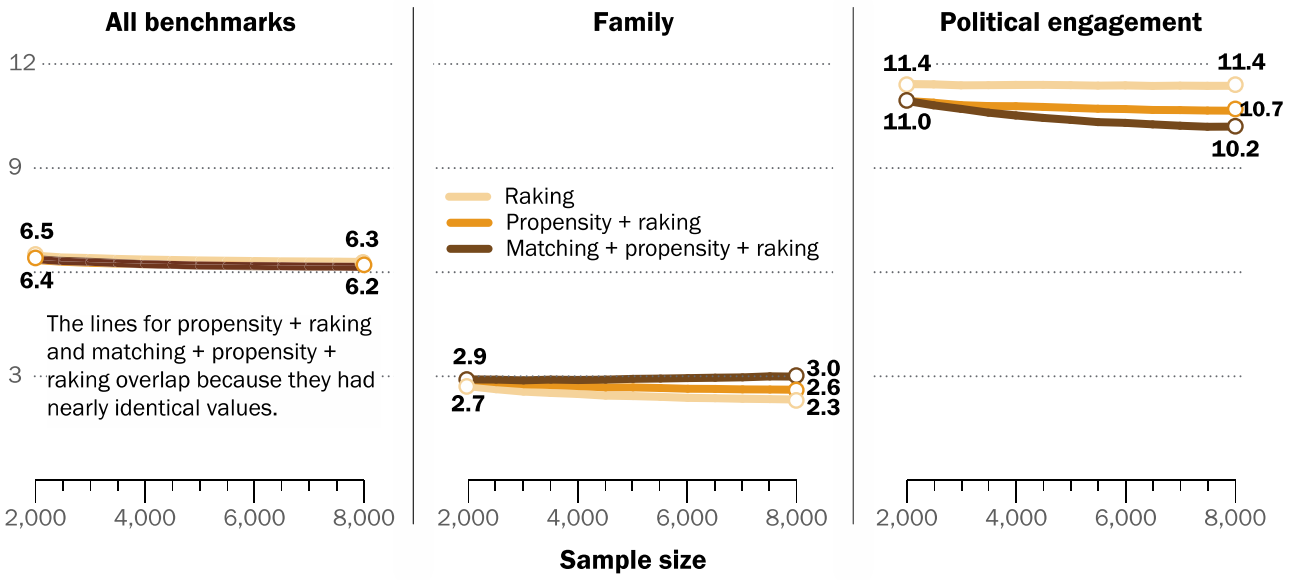
For full sample estimates, the benefits of complex statistical methods are situational

If adjustment usually involves a trade-off between reducing systematic error (bias) but increasing random error (variance), what is the best approach? To find the right balance between bias reduction and increased variability, statisticians often use a measure known as *root mean squared error* (RMSE). RMSE measures the combined effect of both bias *and* variance on the total amount of error in a survey estimate. Although methods that combine matching with other techniques appear to have a slight edge when it comes to bias reduction, the fact that they also tend to have a larger margin of error means that any gains in accuracy may be overwhelmed by large increases in variance.

To test this, the average RMSE was computed for all 24 benchmark variables and compared across three adjustment methods: raking, because it is most common in practice; the two-stage P+R, which produced slightly less biased estimates than raking on its own with the same margin of error; and the three-stage M+P+R technique, which generally had the lowest estimated bias at the expense of discarding interviews. For brevity, the discussion is restricted to the cases where both demographic and political variables are used, although the general pattern is the same.

When bias is high, complex methods help more than they hurt, but the opposite is true when bias is low

Average root mean squared error (RMSE) across 24 benchmarks (percentage points)



Note: Figures are based on adjustments performed using both demographic and political variables.
Source: Pew Research Center analysis of three online opt-in surveys.
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

The study found that, averaged over all 24 benchmark variables, P+R and M+P+R are indistinguishable from one another at every sample size – both having an average RMSE between 6.4 and 6.2 percentage points at sizes n=2,000 and n=8,000 respectively. Simply put, in the aggregate, the total amount of error was the same for both methods. On average, M+P+R produced estimates with slightly more variability than P+R, but made up for it through lower bias. Raking was only slightly higher, going from 6.5 at n=2,000 to 6.3 at n=8,000 – a difference of only 0.2 points.

Although these methods were all roughly equivalent in the aggregate, there were important differences for some survey topics depending on their level of bias prior to adjustment. For most topics, the pattern was consistent with what we saw across all variables. However, for two topics in particular, a different pattern emerged. For questions related to family, raking produced the lowest RMSE, followed by P+R, with M+P+R appreciably higher than the others. Before weighting, the family-related variables had the lowest average bias out of all of the topics, and weighting had little effect. Consequently, none of these estimates saw much in the way of bias reduction, no matter what method was used. With raking and P+R, there is at least the benefit of

lower variance at larger sample sizes, although P+R does slightly worse due to its greater complexity. With M+P+R, the discarded interviews are largely wasted, because there is no bias reduction to offset the greater variability.

The opposite is true for political engagement, which had the highest estimated bias prior to adjustment. Here, the gains from more effective matching at larger starting sample sizes, even after discarding 6,500 out of 8,000 interviews, outweighed the benefits of lower variability that come with methods that use the full sample.

Altogether, these findings suggest that the greater efficacy of complex statistical techniques is highly situational. The three-stage M+P+R method produced real improvements in the total error for the political engagement benchmarks, even accounting for a substantial penalty in terms of variability. Even so, the estimated bias for these measures was high to begin with, and even the most effective adjustment left a great deal of bias remaining. When bias is low, the added complexity simply increases the total level of error relative to simpler methods, as was the case for the benchmarks related to family composition. For most other topics the differences were minimal.

Acknowledgements

This report was made possible by The Pew Charitable Trusts. Pew Research Center is a subsidiary of The Pew Charitable Trusts, its primary funder.

This report is a collaborative effort based on the input and analysis of the following individuals:

Research team

Andrew Mercer, *Senior Research Methodologist*
Arnold Lau, *Research Analyst*
Courtney Kennedy, *Director, Survey Research*
Nick Hatley *Research Analyst*
Scott Keeter, *Senior Survey Advisor*

Communications and editorial

Rachel Weisel, *Communications Manager*
Hannah Klein, *Communications Associate*
David Kent, *Copy Editor*
Travis Mitchell, *Digital Producer*

Graphic design and web publishing

Bill Webster, *Information Graphics Designer*

Colleagues both within and outside of Pew Research Center contributed greatly to the development and execution of this study. We would especially like to thank Claudia Deane for her editorial contributions, as well as Trent Buskirk for his helpful guidance and advice.

Appendix A: Survey methodology

This study, sponsored by Pew Research Center, used online opt-in survey data collected by three commercial vendors. Each vendor provided their own sample and administered the survey themselves, based on a common questionnaire.

While no vendor makes the claim that that these are probability-based samples, they are intended to represent the noninstitutionalized U.S. adult population, though individuals who do not have internet access cannot join these panels and are as such not covered. The precise methods used to recruit the members of each panel are proprietary, and as a result, additional coverage properties are unknown.

Each vendor was provided with the same set of demographic quota targets based on estimates from the 2014 American Community Survey, as well as a measure of population density based on the 2010 decennial census, although differences capabilities and procedures meant that they were implemented slightly differently by each vendor. Only one vendor (Vendor 2) was able to quota on population density. Field dates and implementation details for each vendor are provided below:

Vendor 1 – June 15-28, 2016, n=10,606

Vendor 1 used sample quotas for sex by age, sex by education, age by education, census region and race and Hispanic ethnicity. Respondents sampled by Vendor 1 could opt to take the survey in English or in Spanish.

Vendor 2 – June 17-July 6, 2016, n=10,010

Vendor 2 used sample quotas for sex, age, education, census region, race and Hispanic ethnicity, and population density. Respondents sampled by Vendor 2 could opt to take the survey in English or in Spanish.

Vendor 3 – June 20-June 25, 2016, n=11,247

Vendor 3 used sample quotas on sex by age, sex by education, age by education, census region and race and Hispanic ethnicity. Spanish-language interviews (n=1,518) came from dedicated Spanish-language panels belonging to the vendor, while the rest belonged to English-language panels.

Appendix B: Synthetic population dataset

Several of the adjustment approaches used in this study require a dataset that is highly representative of the U.S. adult population. This dataset essentially serves as a reference for making the survey at hand (e.g., the online opt-in samples) more representative. When selecting a population dataset, researchers typically use a large, federal benchmark dataset such as the American Community Survey (ACS) or Current Population Survey (CPS), as those surveys have high response rates, high population coverage rates and rigorous probability-based sample designs.

One limitation of using a single survey, such as the ACS, is that the only variables that can be used in adjustment are those measured in the ACS. This means that a researcher could adjust on characteristics like age, income and education but not political party affiliation, religious affiliation or voter registration. One solution is to take several benchmark datasets measuring somewhat different variables and combine them to create a *synthetic* population dataset.[25] Questions that the ACS has in common with other benchmark surveys are used to statistically model likely responses to questions that were not asked on the ACS. The subsequent sections detail how the synthetic population dataset was constructed for this study.

Construction of the synthetic population dataset

The synthetic population dataset was constructed in three main steps:

Researchers downloaded public use datasets for nine benchmark surveys and then recoded common variables (e.g., age and education) to be consistent across the surveys. They then rescaled each survey's weights to sum to the nominal sample size.

Each dataset was then sorted according to each record's weight, and divided into 20 strata based on the cumulative sum of the survey weights so that each stratum represented 5% of the total population. Next, a sample of 1,000 cases (interviews) was randomly selected from each stratum with replacement and with probability proportional to the case's weight. This had the effect of "undoing" the weights and producing a 20,000-case dataset for each survey that was representative of the total population.

These 20,000 case datasets were then combined into a single large dataset. Using that combined dataset, researchers produced 25 multiply imputed datasets via the chained equations approach.

[25] Douglas Rivers and Delia Bailey. 2009. "Inference from Matched Samples in the 2008 U.S. National Elections." Presented at the American Statistical Association Joint Statistical Meetings.

After imputation, only the 20,000 cases that originated from the ACS were kept, and all others were discarded. This was done to ensure that the distribution of the main demographic variables precisely matched the ACS distribution, while the imputed variables reflect the distribution that would be expected based on the ACS demographic profile.

Each of these steps is discussed in detail below.

Dataset selection and recoding

Nine datasets were used to construct the synthetic population dataset: the 2015 ACS, the 2015 CPS Annual Social and Economic Supplement (CPS ASEC), the 2013 CPS Civic Engagement Supplement (CPS CivEng), the 2015 CPS Computer and Internet Use Supplement (CPS Internet), the 2015 CPS Volunteer Supplement (CPS Volunteer), the 2014 CPS Voting and Registration Supplement (CPS Voting), the 2014 General Social Survey (GSS), the 2014 Pew Research Center Religious Landscape Study (RLS) and the 2014 Pew Research Center Political Polarization and Typology Survey (Pol.). Each survey contributed a number of variables to the frame. In all, the frame contains 37 variables, with many of these variables present in multiple surveys.

Dataset sample sizes

Dataset	Sample size
ACS	2,424,694
CPS ASEC	144,279
CPS CivEng	27,566
CPS Internet	9,194
CPS Volunteer	80,075
CPS Voting	89,063
GSS	3,842
RLS	35,071
Polarization	10,013

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Rates of missing data for variables used in the synthetic frame

	ACS	CPS ASEC	CPS CivEng	CPS Internet	CPS Vol.	CPS Voting	GSS	RLS	Pol.
Sex	-	-	-	-	-	-	-	-	-
Age	-	-	-	-	-	-	1	2	2
Race/ethnicity	-	-	-	-	-	-	1	2	2
Education	-	-	-	-	-	-	0	1	0
Census division	-	-	-	-	-	-	-	-	-
Marital status	-	-	-	-	-	-	0	1	1
Household size	-	-	-	-	-	-	-	1	2
Number of children	-	-	-	-	-	-	1	0	23
U.S. citizenship	-	-	-	-	-	-	0	0	0
Born in the U.S.	-	-	-	-	-	-	-	1	9
Family income	-	-	-	-	8	-	8	13	10
Employment status	-	-	-	-	-	-	0	NA	68
Employment sector	-	-	-	-	-	-	-	NA	NA
Hours worked per week	-	4	5	4	4	4	0	NA	NA
# of hours worked per week varies	NA	-	-	-	-	-	NA	NA	NA
Military status	-	0	-	-	-	-	0	NA	NA
Home ownership	3	-	-	-	-	-	NA	NA	NA
Metropolitan residence	NA	-	-	-	-	-	NA	NA	NA
Household internet access	-	NA	NA	-	NA	NA	45	NA	NA
Food stamp recipient	-	-	NA	NA	NA	NA	NA	NA	NA
Lived in house or apartment one year ago	-	-	NA	NA	NA	NA	NA	NA	NA
Contacted public official	NA	NA	4	NA	NA	NA	68	NA	NA
Boycotted a company	NA	NA	4	NA	NA	NA	69	NA	NA
Participates in a community group	NA	NA	4	NA	NA	NA	NA	NA	NA
Talks with neighbors	NA	NA	6	NA	NA	NA	NA	NA	NA
Trusts people in neighborhood	NA	NA	7	NA	NA	NA	NA	NA	NA
Household has a tablet or e-book reader	NA	NA	NA	-	NA	NA	NA	NA	NA
Texting or instant messaging	NA	NA	NA	-	NA	NA	NA	NA	NA
Social networking	NA	NA	NA	-	NA	NA	NA	NA	NA
Volunteered	NA	NA	NA	NA	0	NA	NA	NA	NA
Registered to vote in 2014	NA	NA	NA	NA	NA	2	NA	NA	NA
Voted in 2014	NA	NA	NA	NA	NA	2	NA	NA	NA
Party identification	NA	NA	NA	NA	NA	NA	1	-	-
Religion	NA	NA	NA	NA	NA	NA	NA	-	1
Political ideology	NA	NA	NA	NA	NA	NA	3	6	4
Follows government and public affairs	NA	NA	NA	NA	NA	NA	NA	NA	0
Gun ownership	NA	NA	NA	NA	NA	NA	36	NA	NA

Note: Estimates rounded to nearest integer. Dashes indicate no missing data. NA indicates that the variable was not asked in that survey.
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

All nine datasets featured a number of common demographic variables such as sex, age, race and Hispanic ethnicity, education, census division, marital status, household size, number of children, U.S. birth, citizenship status and family income. Other variables were only measured in a subset of the surveys. Volunteering, for example, is present only in the CPS Volunteer Supplement, while party identification is only present in the GSS, the RLS and Pew Research Center's Polarization Survey, none of which are federal government surveys.

Variables that were measured or coded differently across surveys were recoded to be as comparable as possible. This often meant that variables were coarsened. For example, the CPS top-codes age at 85 years or more, so the same coding scheme was applied to all of the other surveys as well. In other cases, this involved treating inconsistent values as missing. For instance, both the ACS and the various CPS surveys ask respondents how many hours they usually work per week. However, the CPS surveys also allow respondents to indicate that the number of hours they usually work per week varies, while the ACS does not have this option. In the above table, missing data for hours worked per week across the CPS surveys is not truly missing; rather, it consists of people who indicated that their hours vary. However, these data are treated as missing for consistency with the way it is asked in the ACS. Imputed values can be interpreted as predicting how those individuals would have answered if they had been asked the ACS question instead.

Stratified sampling

The benchmark datasets differed in sample design and sample sizes. In order to address these differences, we selected exactly 20,000 observations per dataset before appending them together. Sampling was done with replacement and with probability proportional to the case's weight. The sample size was selected in order to provide enough data for the adjustment methods used while still being computationally tractable. For the CPS Internet Supplement, the GSS and the Polarization Survey, this guaranteed that observations would be sampled multiple times.

We used the relevant weights for each dataset. The person-level weight was used for the ACS, the person supplement weight for the CPS ASEC and the self-response supplement weight for the CPS Civic Engagement supplement. The CPS Internet Supplement was filtered down to respondents who had a random respondent weight, because the texting and social networking variables were only measured for these respondents. The nonresponse weight was used for the CPS Volunteer Supplement, while the nonresponse weight accounting for both cross-section and panel cases was used for the GSS. Full sample weights were used for the RLS and the Polarization Survey. Finally, for the CPS Voting Supplement, the second-stage weights were adjusted as recommended by Hur

and Achen[26] to correct for bias resulting from item nonresponse being treated as not having voted. Each of these weights was rescaled to sum to the sample size of each of their respective datasets.

To ensure that the samples contained the correct proportion of cases with both large and small weights, each dataset was sorted according to the weights, and divided into 20 strata, each of which represented 5% of the weighted sample.

Imputation

The nine datasets were then combined into a single dataset, and all missing values were imputed via a "chained equations" approach that iterates through modeling each variable as a function of all the others.[27] For example, if age, sex and education were the only variables, a chained equations approach might first impute age based on sex and education, then sex based on age and education, then education based on age and sex, and would repeat this cycle for some number of iterations in order to achieve stability. This entire procedure is also repeated 25 times, independently of one another, to produce multiple synthetic frames that can be compared against one another to assess variance stemming from the imputation process. Each frame went through 10 iterations.

There are a wide variety of models that can be used to impute each individual variable dependent on all the others, such as regression models or "hot-deck" methods where each missing value is replaced by an observed response from a "similar" unit. For the synthetic population dataset, each variable was imputed using a random forest "hot-deck" method.[28]

After imputation, the final synthetic population dataset was created by deleting all but the cases that were originally from the ACS. This ensures that the demographic distribution closely matches that of the original ACS, while the imputed variables reflect the joint distribution that would be expected based on the variables that each dataset had in common.

Evaluating the imputation quality

We took several steps to ensure that the imputation procedure produced results that accurately reflected the original datasets. First, we crossed each of the imputed variables (e.g., voter registration and party identification) with the fully observed variables (e.g., age, sex and

[26] See Hur, Aram, and Christopher H. Achen. 2013. "Coding Voter Turnout Responses in the Current Population Survey." Public Opinion Quarterly 77 (4), 985-993.
[27] See Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?: Multiple Imputation by Chained Equations." International Journal of Methods in Psychiatric Research 20 (1), 40-49.
[28] For complete details on the random forest imputation procedure that was used in this study, see Doove, L.L., S. Van Buuren, and E. Dusseldorp. 2014. "Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects." Computational Statistics & Data Analysis 72 (April), 92-104.

education), and for each cell, compared the size of the cell in the ACS dataset to its size in the original dataset from which it was imputed. Overall, the imputed distributions were quite close to the originals. The average absolute difference between the imputed and original values for each cross-classification was 2 percentage points. This means that on average, the imputed values not only matched the distribution for the full population, but also matched the distribution within demographic subgroups.

Although the multiple imputation procedure created 25 versions of the synthetic population dataset, only one of them was used to perform the adjustments in this study. One concern with this approach is the possibility that the results could vary widely depending on which of the 25 synthetic populations was used. Although it was not computationally feasible to repeat the entire analysis on each of the imputed datasets, we did repeat one of the adjustment procedures across all 25 datasets in order to assess the degree to which the imputation procedure may be affecting the study's findings.

For each of the 25 imputed datasets, we performed raking with both the demographic and political variables on 1,000 bootstrap samples of n=3,500 following the same procedure that was used in the body of this report. For each substantive category in the 24 benchmark variables, we calculated the weighted percentage for each bootstrapped sample. Then we calculated the *total variance* (mean squared error) for each estimate with all 25,000 bootstrap samples combined. Finally, we calculated the variance for each of the 25 sets of estimates separately and took the average. This is the *within-imputation variance*. This process was repeated for all three vendors.

If the total variance is much larger than the within-imputation variance, then estimated variability and margins of error that use only a single imputation (as was done in this study) would be underestimated. In this case, the total variance was only 1.002 times as large as the average within-imputation variance. This means that the estimated variability described in the report is for all practical purposes the same as if the analysis had been repeated for all 25 imputations.

The reason the two are so close is likely due to the fact the imputation only affects the variability of the survey estimates indirectly, and makes up only a small portion of the survey variability. If we were to compare the total and within-imputation variability for the imputed values themselves (as we might if the synthetic population dataset were the main focus of the analysis rather than simply an input to the weighting), the difference would likely be larger.

Adjustment variables used in the study

The core demographic adjustment variables used in the study were 6-category age, sex, 5-category educational attainment, race and Hispanic ethnicity, and census division. The expanded political variables add to this 3-category political party affiliation, 3-category political ideology, voter registration, and whether the respondent identifies as an evangelical Christian.

The following table compares the distribution of the adjustment variables on the synthetic population dataset versus from one of the original high-quality survey datasets used to create the synthetic dataset. All demographic variables were fully observed on the ACS, so the synthetic frame will differ from the original source only on the set of expanded political variables.

The largest difference between the source survey and the synthetic frame was on political ideology. The estimated share of self-described conservatives was 32% in the GSS versus 35% in the synthetic frame. The latter estimate is similar to measures from Pew Research Center's Religious Landscape Study and the Political Polarization and Typology Survey, which were also used in the frame. The exact reason for this discrepancy is unclear, but there are several potential factors. Unlike the Center's measures, which are collected via live telephone interviewing, the GSS question is administered in-person using a showcard. In addition, the GSS question uses a seven-point scale, while the Center's questions use a five-point scale. Finally, there may be important differences between the demographic makeup of respondents to the GSS and respondents to the ACS.

Distribution of adjustment variables from original source and from synthetic population dataset

Demographic variables

Adjustment variable	Source	Response category	Source estimate (%)	Synthetic dataset estimate (%)	Notes
Sex	American Community Survey (2015)	Male	48	48	
		Female	52	52	
Education	American Community Survey (2015)	Less than HS	13	13	
		HS graduate	28	28	
		Some college	31	32	
		College graduate	18	18	
		Postgraduate	10	10	
Race and Hispanic ethnicity	American Community Survey (2015)	White, non-Hispanic	65	65	
		Black, non-Hispanic	12	12	
		Hispanic	16	15	
		Asian	6	6	
		Other race	3	3	
Age	American Community Survey (2015)	18-24	13	13	.
		25-34	18	18	
		35-44	17	16	
		45-54	17	18	
		55-64	17	16	
		65+	19	19	

Note: Source estimates are weighted and rounded to the nearest integer.
"For Weighting Online Opt-In Samples, What Matters Most?"

Distribution of adjustment variables from original source and from synthetic frame (continued)

Demographic variables (continued)

Adjustment variable	Source	Response category	Source estimate (%)	Synthetic frame estimate (%)	Notes
Census division	American Community Survey (2015)	East North Central	15	15	
		East South Central	6	6	
		Middle Atlantic	13	13	
		Mountain	7	7	
		New England	5	5	
		Pacific	16	16	
		South Atlantic	20	20	
		West North Central	7	6	
		West South Central	12	11	

Political variables

Adjustment variable	Source	Response category	Source estimate (%)	Synthetic frame estimate (%)	Notes
Party	General Social Survey (2014)	Republican	22	22	Source estimate does not add up to 100% due to item nonresponse.
		Independent/Other	46	48	
		Democrat	32	30	
Ideology	General Social Survey (2014)	Liberal	27	26	Source estimate does not add up to 100% due to item nonresponse.
		Moderate	38	39	
		Conservative	32	35	
Is an evangelical Christian	Pew Research Center Religious Landscape Study (2014)	Yes	29	27	
		No	71	73	
Is registered to vote	CPS Voting and Registration Supplement (Nov 2014)	Yes	65	66	Source estimate does not add up to 100% due to item nonresponse.
		No	32	34	

Note: Source estimates are weighted and rounded to the nearest integer.
Source: “For Weighting Online Opt-In Samples, What Matters Most?”

Appendix C: Adjustment procedures

Raking

Raked weights were created using the marginal distributions of the adjustment variables as derived from the synthetic population dataset, along with all two-way interactions of collapsed versions of the demographic variables. For the interactions, the 18-24 and 25-34 age categories were combined, the less than high school and high school graduate categories were combined, race and Hispanic ethnicity was collapsed into white vs nonwhite, and census region was used instead of census division. This was done to avoid low adjustment cell counts and the chance that a subsample would yield a cell with no observations.

The **calibrate** function in the **survey** package in R[29] was used for raking. When raking was used as the final step in a combination procedure, such as matching followed by propensity weighting followed by raking, the interim weights were trimmed at the 5th and 95th percentiles. No trimming was applied to the final weights.

Random forest

Both the matching and propensity adjustments used in this study were carried out using a statistical approach called random forest. Random forest models belong to a more general set of machine-learning models called classification and regression trees. These models work by partitioning the data into smaller and smaller subsets, called "nodes." The partitions resemble a tree-like structure, hence the name. The further down the tree, the more all observations within a node agree with each other over whatever the outcome measure is.

The covariates fed into the model become the basis for which the data is split into nodes. For instance, one such split early on may divide the data into a node for the male cases and another node for the female cases. Each of those nodes may then be split further on some other covariate. The tree is considered fully grown when either all observations within every single node agree on the outcome, or when any further splitting would bring the number of observations in a node below a user-defined minimum size. The nodes at the end of the tree are called *terminal nodes*.

In random forest models, numerous trees are grown, with each tree being fit on a bootstrapped sample of the data, and with each tree's partitions being determined using only a subset of the covariates in the full model. Predicted probabilities and proximity measures are then calculated by averaging across all the trees.

[29] Thomas Lumley. 2017. "survey: Analysis of Complex Survey Samples." R package version 3.32.

For this study, random forest models were fit in R using the **ranger** package.[30] All models used 1,000 trees and had a minimum node size of 100.

Propensity weighting

The online opt-in sample and the full synthetic population dataset were combined and a new binary variable was created with a value of 1 if the case came from the synthetic dataset, and zero otherwise. A random forest model was then fit with the binary variable as the outcome and the adjustment variables as the covariates. The model then returned a predicted probability p that each case in the combined dataset came from the synthetic dataset. The quantity $1 - p$ is then the predicted probability that each case in the combined dataset came from the online opt-in sample. Subsequently, for each case in the online opt-in sample, the propensity weight was $\frac{p}{1-p}$.

The resulting weights were rescaled to sum to the size of the online opt-in sample.

Matching

For matching, the online opt-in sample was combined with a target sample of 1,500 cases that were randomly selected from the synthetic population. A random forest model was then used to predict whether or not each case belonged to the target sample based on the adjustment variables. The models used 1,000 trees and had a minimum node size of 100.

Once the model was fit, the "distance" between each case in the target sample and all of the cases in the survey sample was calculated. For a given tree, cases that are similar to one another end up in the same terminal node. The *random forest proximity* between any two cases is simply the number of trees in which they were placed in the same node divided by the total number of trees used in the model. For example, if a particular pair of cases ended up in the same terminal node in 300 trees and in different terminal nodes in the 700 other trees, then the random forest proximity for that pair would be 0.3. A proximity close to 1 means the cases are very similar to one another, while a proximity close to zero means they are very different. The **RcppEigen** package[31] was used to speed up calculation.

After the random forest proximity for each pair was calculated, both the synthetic dataset and the online opt-in sample were sorted in random order. The final matched sample was selected by sequentially matching each of the 1,500 cases in the synthetic frame sample to the case from the online opt-in sample with which it has the largest random forest proximity, with ties being broken

[30] Marvin N. Wright and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." Journal of Statistical Software 77(1), 1-17.
[31] Douglas Bates and Dirk Eddelbuettel. 2013. "Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package." Journal of Statistical Software 52(5), 1-24.

randomly. Matched cases were given weights of 1, while unmatched cases were given weights of zero.

Appendix D: Sources and details for benchmarks

Topic: Civic engagement

Benchmark item	Source	Question text	Response category	Benchmark estimate (%)	Notes
Talked with neighbors	CPS Civic Engagement Supplement (Nov 2013)	During a typical month in the past year, how often did you talk with any of your neighbors?	Basically every day	12.1	
			A few times a week	28.9	
			A few times a month	21.6	
			Once or less than once a month	19.5	
			Not at all	12.3	
Trusts neighbors	CPS Civic Engagement Supplement (Nov 2013)	How much do you trust the people in your neighborhood? In general, do you trust …	All of the people in your neighborhood	13.7	
			Most of the people in your neighborhood	37.3	
			Some of the people in your neighborhood	33.2	
			None of the people in your neighborhood	8.6	
Participated in a school group, neighborhood, or community association	CPS Civic Engagement Supplement (Nov 2013)	In the last 12 months, that is since June 2015, have you participated in a school group, neighborhood, or community association such as PTA or neighborhood watch group?	Yes	13.7	
			No	82.1	
Volunteered	CPS Volunteer Supplement (Sep 2015)	In the last 12 months, that is since June of last year, have you done any volunteer activities through or for an organization? Sometimes people don't think of activities they do infrequently or activities they do for children's schools or youth organizations as volunteer activities. Since June of last year, have you done any of these types of volunteer activities?	Yes	24.8	The variable used to produce this estimate is a recode of two Yes/No questions from the CPS. The second question clarifies the definition of 'volunteer activities' and is asked if respondents skipped or answered no to the first question.
			No	75.0	

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Topic: Financial

Benchmark item	Source	Question text	Response category	Benchmark estimate (%)	Notes
Employment status	General Social Survey (2016)	Last week, were you working full time, part time, going to school, keeping house, or what?	Working full time	47.2	
			Working part time	13.2	
			With a job, but not at work because of temporary illness, vacation, strike	1.9	
			Unemployed, laid off, looking for work	4.2	
			Retired	17.0	
			In school	3.2	
			Keeping house	10.3	
Home ownership	American Community Survey (2015)	Is your house, apartment, or mobile home ...	Owned by you or someone in this household with a mortgage or loan.	43.1	On the ACS, this question was not asked of people who lived in non-institutional group quarters (such as dormitories).
			Owned by you or someone in this household free and clear	22.2	
			Rented	31.4	
			Occupied without payment of rent	1.6	
Family income	CPS Annual Social and Economic Supplement (Mar 2016)	Which category represents the total combined income of all members of your FAMILY during the past 12 months? This includes money from jobs, net income from business, farm or rent, pensions, dividends, interest, social security payments and any other money income received by members of your family who are 15 years of age or older.	Less than $5,000	2.6	
			$5,000 to $7,499	1.4	
			$7,500 to $9,999	1.9	
			$10,000 to $12,499	2.5	
			$12,500 to $14,999	2.5	
			$15,000 to S19,999	3.9	
			$20,000 to $24,999	5.1	
			$25,000 to $29,999	5.4	
			$30,000 to $34,999	5.5	
			$35,000 to $39,999	5.1	
			$40,000 to $49,999	8.6	
			$50,000 to $59,999	8.3	
			$60,000 to $74,999	10.4	
			$75,000 to $99,999	12.5	
			$100,000 to $149,999	13.0	
			$150,000 to more	11.2	

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Topic: Financial (continued)

Benchmark item	Source	Question text	Response category	Benchmark estimate (%)	Notes
Food stamps	CPS Annual Social and Economic Supplement (Mar 2016)	Did anyone in your household get food stamps or use a food stamp benefit card at any time during 2015? *Do not include WIC benefits.*	Yes	10.6	
			No	89.4	
Health insurance	National Health Interview Survey (2015)	Are you covered by any kind of health insurance or some other kind of health care plan? Include health insurance obtained through employment or purchased directly as well as government programs like Medicare and Medicaid that provide medical care or help pay medical bills.	Yes	89.0	
			No	10.4	

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Topic: Family

Benchmark item	Source	Question text	Response category	Benchmark estimate (%)	Notes
Marital status	American Community Survey (2015)	What is your marital status?	Now married	50.5	
			Widowed	5.9	
			Divorced	11.5	
			Separated	2.1	
			Never married	30.0	
Children in household	American Community Survey (2015)	And how many children younger than 18 years of age live in your household?	No children	65.0	This figure is calculated by counting the number of children under 18 in each ACS household.
			One or more children	35.0	
Household size	American Community Survey (2015)	N/A	1	15.2	This figure is calculated by adding the number of adults in each ACS household to the number of children under 18 in each ACS household.
			2	32.9	
			3+	51.9	

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Topic: Personal

Benchmark item	Source	Question text	Response category	Benchmark estimate (%)	Notes
Lived in house or apartment one year ago	American Community Survey (2015)	Did you live in your house or apartment one year ago?	Same house	85.7	
			Different house in US	13.6	
			Different house outside US	0.7	
Active duty military service	American Community Survey (2015)	Have you ever served on active duty in the U.S. Armed Forces, Reserves, or National Guard?	Have been on active duty	8.0	The variable used to produce this estimate is a recode that collapses people who are currently on active duty and people who were on active duty in the past, and does not consider Reserves or National Guard as active duty.
			Have never been on active duty	92.0	
U.S. citizenship	American Community Survey (2015)	Are you a citizen of the United States?	Yes, a U.S. citizen	91.6	
			No, not a U.S. citizen	8.4	
Gun ownership	General Social Survey (2016)	Do you happen to have in your home or garage any guns or revolvers?	Yes	31.7	
			No	65.4	
Smoking	National Health Interview Survey (2015)	Have you smoked at least 100 cigarettes in your ENTIRE LIFE? Do you NOW smoke cigarettes every day, some days, or not at all?	Smoke every day	11.4	The variable used to produce this estimate collapses two questions from the NHIS.
			Smoke some days	3.7	
			No longer smoke	21.8	
			Have never smoked	62.8	
Food allergies	National Health and Nutrition Examination Survey (2007)	Do you have any food allergies?	Yes	10.0	The NHANES 2007 was used due to this question not having been asked in NHANES 2016.
			No	89.8	

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Topic: Political engagement

Benchmark item	Source	Question text	Response category	Benchmark estimate (%)	Notes
Voted in 2012	CPS Voting and Registration Supplement (Nov 2012)	In the 2012 presidential election between Barack Obama and Mitt Romney, did things come up that kept you from voting, or did you happen to vote?	Voted	50.2	These estimates use the adjustment recommended in Hur and Achen (2013) to correct for bias resulting from the fact that item nonrespondents are treated as not having voted in the CPS. Adjustment factors for 2012 can be found at: http://www.electproject.org/home/voter-turnout/cps-methodology These estimates are further adjusted to approximate the percentage of adults in 2016 who voted in 2012. The adjustment was done by using the ACS to break out the total adult population in 2016 by citizenship, age group and race. Each break was then multiplied by the probability that said group voted 4 years ago (in 2012), obtained from the CPS. Finally, the breaks were added together to get estimates of voting in 2012 for the total 2016 adult population.
			Did not vote (includes too young to vote)	49.8	
Voted in 2014	CPS Voting and Registration Supplement (Nov 2014)	In the 2014 midterm election, did things come up that kept you from voting, or did you happen to vote?	Voted	32.7	These estimates are adjusted to correct for item nonresponse bias and to approximate the percentage of adults in 2016 who voted in 2014, as described in the notes for the 'Voted in 2012' benchmark estimate.
			Did not vote (includes too young to vote)	67.3	
Contacted or visited a public official	CPS Civic Engagement Supplement (Nov 2013)	In the past 12 months, that is since June 2015, have you contacted or visited a public official—at any level of government—to express your opinion?	Yes	11.2	
			No	85.1	

"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Topic: Technology

Benchmark item	Source	Question text	Response category	Benchmark estimate (%)	Notes
Tablet use	CPS Computer and Internet Use Supplement (July 2015)	Do you use a tablet or e-book reader?	Yes	37.4	
			No	62.6	
Texting or instant messaging	CPS Computer and Internet Use Supplement (July 2015)	What about texting or instant messaging? Do you use a texting or instant messaging service?	Yes	82.4	
			No	17.6	
Social networking	CPS Computer and Internet Use Supplement (July 2015)	What about social networking? Do you use social networks such as Facebook or Twitter?	Yes	67.5	
			No	32.5	

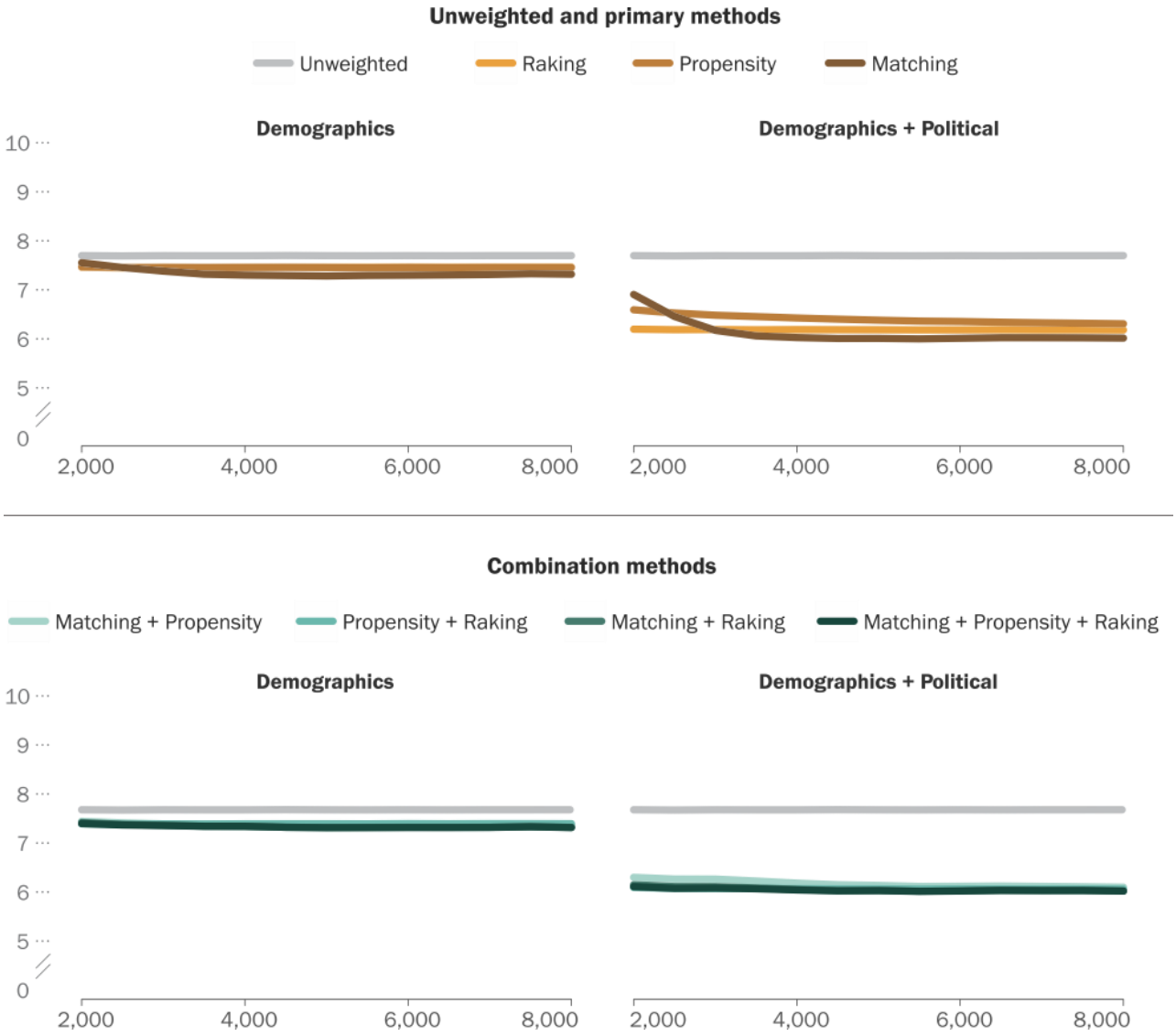
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Appendix E: Average estimated bias by vendor

Vendor 1: Average estimated bias across all weighting procedures

Average absolute differences between population benchmarks and sample estimates (percentage points)

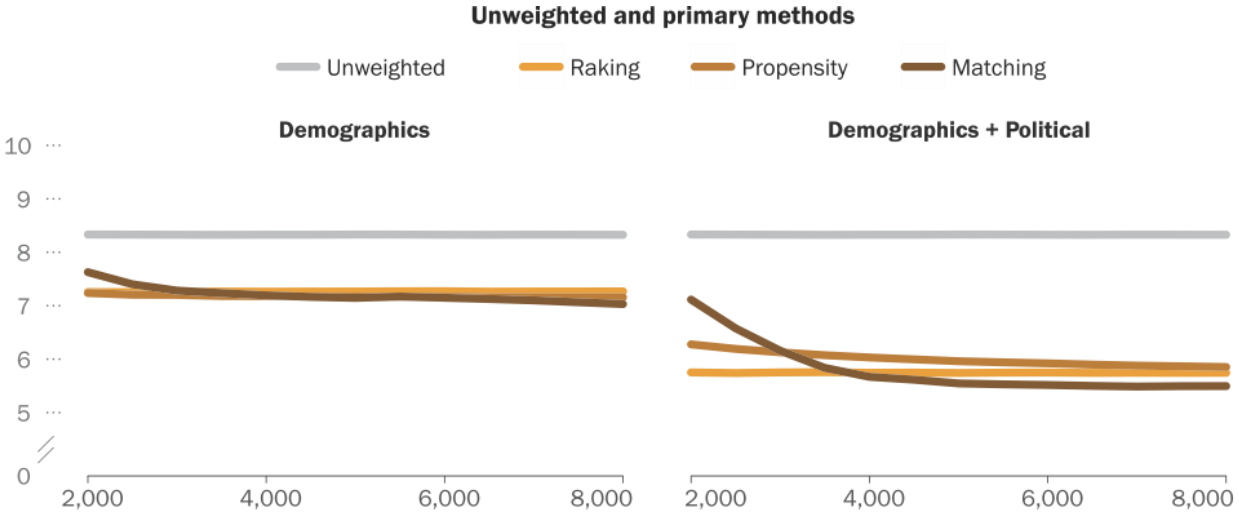


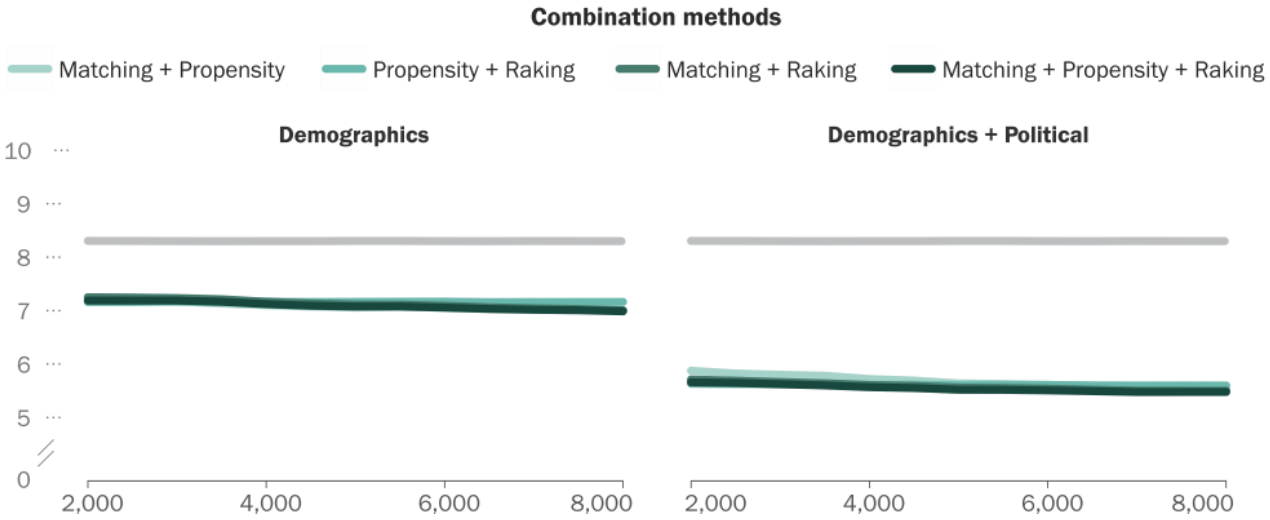
"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Vendor 2: Average estimated bias across all weighting procedures

Average absolute differences between population benchmarks and sample estimates (percentage points)



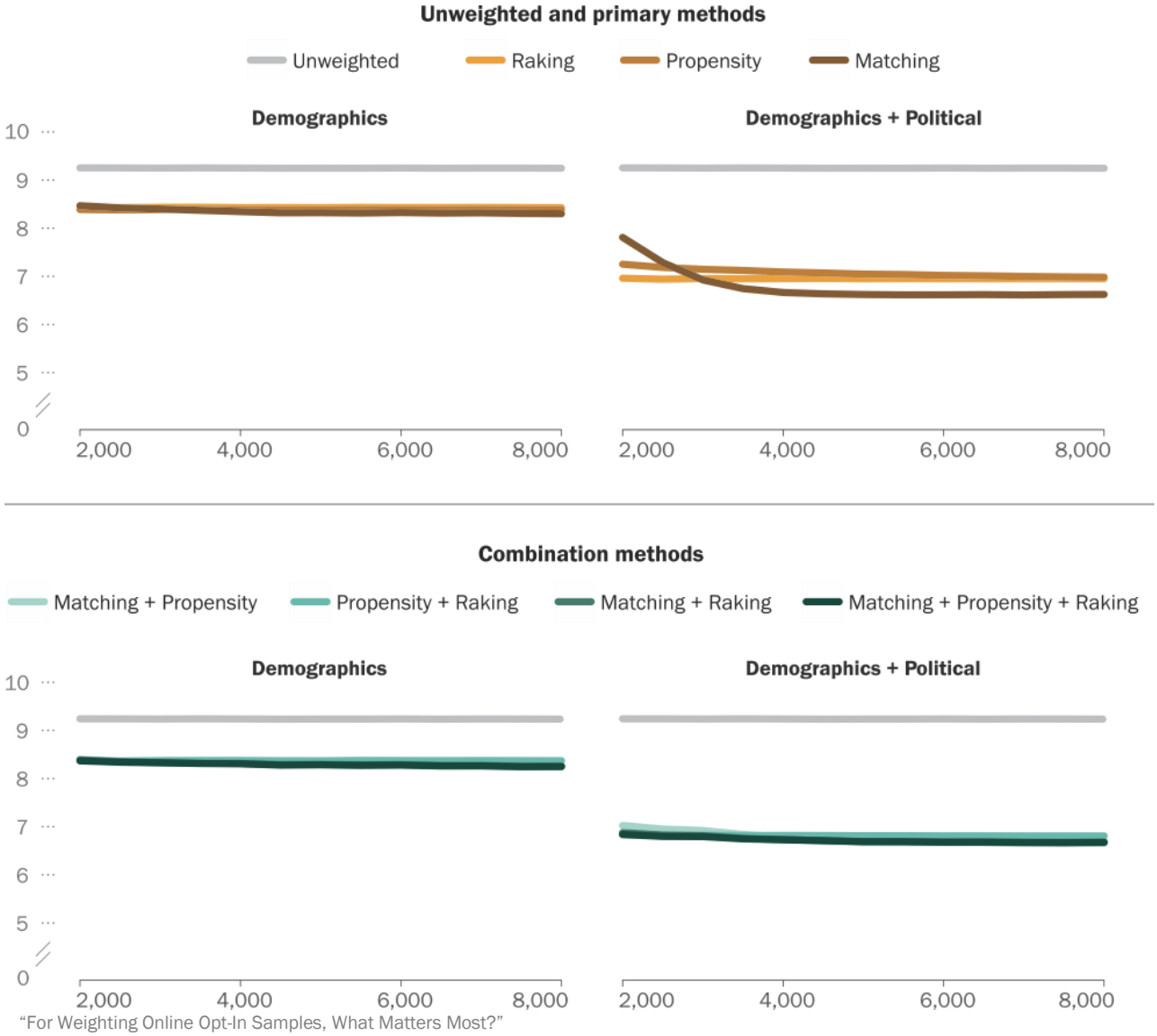


"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Vendor 3: Average estimated bias across all weighting procedures

Average absolute differences between population benchmarks and sample estimates (percentage points)



"For Weighting Online Opt-In Samples, What Matters Most?"

PEW RESEARCH CENTER

Appendix F: Questionnaire

PEW RESEARCH CENTER
ONLINE OPT-IN ADJUSTMENT STUDY
Questionnaire for Programming

[INTRO SCREEN]
Thank you for participating in this survey and we hope you enjoy it. Your answers will be used for research purposes only and will never be attributed to you. The survey should take about 15 minutes for most people to complete.

ASK ALL:
PRESAPP Do you approve or disapprove of the way Barack Obama is handling his job as president?

1 Approve
2 Disapprove

ASK ALL:
HAPPY Generally, how would you say things are these days in your life? Would you say that you are …

1 Very happy
2 Pretty happy
3 Not too happy

ASK ALL:
FOLGOV Would you say you follow what's going on in government and public affairs …

1 Most of the time
2 Some of the time
3 Only now and then
4 Hardly at all

ASK ALL:
VOTEGEN If the 2016 presidential election were being held today, would you vote for …
[RANDOMIZE OPTIONS 1 AND 2 WITH OPTION 3 ALWAYS LAST]

1 Donald Trump, the Republican
2 Hillary Clinton, the Democrat
3 Vote for neither/Other

ASK IF SELECTED CANDIDATE (VOTEGEN=1,2):
VOTEGEN2 And would you say…

[PROGRAMMING NOTE: FILL BASED ON RESPONSE TO VOTEGEN]
1 You are certain to vote for **[**Clinton over Trump/Trump over Clinton**]**
2 There is a chance you might change your mind

ASK IF NEITHER/OTHER CANDIDATE OR SKIPPED VOTEGEN (VOTEGEN=3 OR MISSING):
VOTEGEN3
[IF NEITHER/OTHER IN VOTEGEN DISPLAY:]
And even though you don't plan to support Donald Trump or Hillary Clinton, if you had to choose, would you say you …
[IF SKIPPED VOTEGEN DISPLAY:]
If you had to choose, would you say you …

[RANDOMIZE OPTIONS IN SAME ORDER AS VOTEGEN]

1 Lean more toward Donald Trump
2 Lean more toward Hillary Clinton
3 Neither

[PROGRAMMING NOTE: Display choice 3 Neither only if the question is skipped without selecting choice 1 or 2]

Soft Prompt: If you would not support either candidate please select answer choice Neither. If you would like to skip, click Next.

ASK ALL:
TALK_CPS During a typical month in the past year, how often did you talk with any of your neighbors?

1 Basically every day
2 A few times a week
3 A few times a month
4 Once a month
5 Not at all

ASK ALL:
TRUST_CPS How much do you trust the people in your neighborhood? In general, do you trust …

1 All of the people in your neighborhood
2 Most of the people in your neighborhood
3 Some of the people in your neighborhood
4 None of the people in your neighborhood

ASK ALL:
COMGRP_CPS In the last 12 months, that is since June 2015, have you participated in a school group, neighborhood, or community association such as PTA or neighborhood watch group?

1 Yes
2 No

ASK ALL:
VOL1 We are interested in volunteer activities for which people are not paid, except perhaps expenses. We only want you to include volunteer activities that you did through or for an organization, even if you only did them once in a while. In the last 12 months, that is since June of last year, have you done any volunteer activities through or for an organization?

1 Yes
2 No

ASK IF NO OR SKIPPED IN VOL1 (VOL1 = 1 OR MISSING):
VOL2 Sometimes people don't think of activities they do infrequently or activities they do for children's schools or youth organizations as volunteer activities. Since June of last year, have you done any of these types of volunteer activities?

1 Yes
2 No

ASK ALL:
ACAAPP Do you approve or disapprove of the health care law passed by Barack Obama and Congress in 2010?

1 Approve
2 Disapprove

ASK ALL:
MRJLEGAL Do you think the use of marijuana should be made legal, or not?

1 Yes, legal
2 No, illegal

On another topic…

[PROGRAMMING NOTE: RANDOMIZE ORDER OF DISCRIMA, DISCRIMB, DISCRIMC]
ASK ALL:
DISCRIMA Is there a lot of discrimination against blacks, or not?

1 Yes, there is a lot of discrimination
2 No, not a lot of discrimination

ASK ALL:
DISCRIMB Is there a lot of discrimination against gays and lesbians, or not?

1 Yes, there is a lot of discrimination
2 No, not a lot of discrimination

ASK ALL:
DISCRIMC Is there a lot of discrimination against Hispanics, or not?

1 Yes, there is a lot of discrimination
2 No, not a lot of discrimination

ASK ALL:
FOLNEWS Would you say you follow the news…

1 All or most of the time
2 Some of the time
3 Only now and then
4 Hardly ever

ASK ALL:
NEWSCLOSEA How closely do you follow … International news?

1 Very closely
2 Somewhat closely
3 Not very closely
4 Not at all closely

ASK ALL:
NEWSCLOSEB How closely do you follow … National news?
1 Very closely
2 Somewhat closely
3 Not very closely
4 Not at all closely

ASK ALL:
NEWSCLOSEC How closely do you follow … Local news?
1 Very closely
2 Somewhat closely
3 Not very closely
4 Not at all closely

[PROGRAMMING NOTE: RANDOMIZE ORDER OF PAIR1, PAIR2, PAIR3]
ASK ALL:
PAIR1 Which statement comes closer to your views, even if neither is exactly right?

[RANDOMIZE STATEMENTS]

1 Government should do more to solve problems
2 Government is doing too many things better left to businesses and individuals

---

ASK ALL:
PAIR2 Which statement comes closer to your views, even if neither is exactly right?

[RANDOMIZE STATEMENTS]

1 The economic system in this country unfairly favors powerful interests
2 The economic system in this country is generally fair to most Americans

---

ASK ALL:
PAIR3 Which statement comes closer to your views, even if neither is exactly right?

[RANDOMIZE STATEMENTS]

1 Immigrants today strengthen our country because of their hard work and talents
2 Immigrants today are a burden on our country because they take our jobs, housing and health

---

ASK ALL:
OWNGUN_GSS Do you happen to have in your home or garage any guns or revolvers?
1 Yes
2 No

ASK ALL:
EVSMK_NHIS Have you smoked at least 100 cigarettes in your ENTIRE LIFE?
1 Yes
2 No

ASK IF YES IN EVSMK_NHIS (EVSMK_NHIS=1):
NOWSMK_NHIS Do you NOW smoke cigarettes ...

1 Every day
2 Some days
3 Not at all

ASK ALL:

RACEREL Do you think race relations in the United States are getting better, getting worse or staying about the same?

1 Getting better
2 Getting worse
3 Staying about the same

ASK ALL:

PUB_OFF_CPS In the past 12 months, that is since June 2015, have you contacted or visited a public official—at any level of government—to express your opinion?

1 Yes
2 No

ASK ALL:

PRTYPREF_GSS Generally speaking, do you usually think of yourself as a Republican, Democrat, independent, or what?

1 Republican
2 Democrat
3 Independent
4 Other

ASK IF REPUBLICAN OR DEMOCRAT IN PRTYPREF_GSS (PRTYPREF_GSS=1,2):

PRTYSTRG_GSS Would you call yourself a strong [FILL FROM PRTYPREF_GSS Democrat/Republican] or not a very strong [Democrat/Republican]?

1 Strong
2 Not very strong

ASK IF INDEPENDENT, OTHER OR SKIPPED PRTYPREF_GSS (PRTYPREF_GSS=3,4 OR MISSING):

PRTYIND_GSS Do you think of yourself as closer to the Republican or Democratic Party?

1 Republican
2 Democrat

ASK ALL:

POLVIEWS_GSS We hear a lot of talk these days about liberals and conservatives. On a seven-point scale where the political views that people might hold are arranged from extremely liberal to extremely conservative – where would you place yourself on this scale?

1 Extremely liberal
2 Liberal
3 Slightly liberal
4 Moderate, middle of the road
5 Slightly conservative
6 Conservative
7 Extremely conservative

ASK ALL:
TABLET_CPS Do you use a tablet or e-book reader?
1 Yes
2 No

ASK ALL:
TEXTIM_CPS What about texting or instant messaging? Do you use a texting or instant messaging service?
1 Yes
2 No

ASK ALL:
SOCIAL_CPS What about social networking? Do you use social networks such as Facebook or Twitter?
1 Yes
2 No

ASK ALL:
ADULTS_HH How many adults, ages 18 and older, including yourself, live in your household?

[ENTER NUMBER BETWEEN 1 AND 20]

ASK ALL:
CHILDREN_HH And how many children younger than 18 years of age live in your household? (Please fill in zero "0" if no children)

[ENTER NUMBER BETWEEN 0 AND 20]

ASK ALL:
HOME_ACS Is your house, apartment, or mobile home …

1 Owned by you or someone in this household with a mortgage or loan. *Include home equity loans*
2 Owned by you or someone in this household free and clear (without a mortgage or loan)
3 Rented
4 Occupied without payment of rent

ASK ALL:
TENURE_ACS Did you live in your house or apartment one year ago?
1 Yes, this house
2 No, outside the United States and Puerto Rico
3 No, different house in the United States or Puerto Rico

ASK ALL:
GENDER Are you male or female?

1 Male
2 Female

ASK ALL:
AGE What is your age?

[ENTER NUMBER BETWEEN 18 AND 110]

ASK ALL:
EDUC_ACS What is the highest degree or level of school that you have COMPLETED?

1 No schooling completed
2 Nursery school
3 Kindergarten
4 Grade 1 through 11
5 12th Grade – **NO DIPLOMA**
6 Regular high school diploma
7 GED or alternative credential
8 Some college credit, but less than 1 year of college credit
9 1 or more years of college credit, no degree
10 Associate's degree (for example: AA, AS)
11 Bachelor's degree (for example: BA, BS)
12 Master's degree (for example: MA, MS, MEng, MEd, MSW, MBA)
13 Professional degree beyond a bachelor's degree (for example: MD, DDS, DVM, LLB,JD)
14 Doctorate degree (for example: PhD, EdD)

ASK ALL:
MARITAL_ACS What is your marital status?

1 Now married
2 Widowed
3 Divorced
4 Separated
5 Never married

ASK ALL:
MIL_ACS Have you ever served on active duty in the U.S. Armed Forces, Reserves, or National Guard?

1 Never served in the military
2 Only on active duty for training in the Reserves or National Guard
3 Now on active duty
4 On active duty in the past, but not now

ASK ALL:
HISP_ACS Are you of Hispanic, Latino or Spanish Origin?

1 No, not of Hispanic, Latino or Spanish origin
2 Yes, Mexican, Mexican am., Chicano
3 Yes, Puerto Rican
4 Yes, Cuban
5 Yes, another Hispanic, Latino, or Spanish origin

ASK ALL:
RACE_ACS What is your race? [Choose all that apply]

1 White
2 Black or African Am.
3 Asian
4 American Indian or Alaska Native
5 Native Hawaiian or Pacific Islander
6 Some other race

ASK ALL:
BORN_ACS Where were you born?

1 Inside the United States
2 Outside the United States

ASK IF BORN OUTSIDE THE USE OR SKIPPED BORN_ACS (BORN_ACS=2 OR MISSING):
CITIZEN Are you a citizen of the United States?

1 Yes, a U.S. citizen
2 No, not a U.S. citizen

ASK ALL:
INSURE_NHIS Are you covered by any kind of health insurance or some other kind of health care plan? Include health insurance obtained through employment or purchased directly as well as government programs like Medicare and Medicaid that provide medical care or help pay medical bills.

1 Yes
2 No

ASK ALL:
FDALL_NHANES Do you have any food allergies?

1 Yes
2 No

ASK ALL
RELIG What is your present religion, if any?

1 Protestant (for example, Baptist, Methodist, Non-denominational, Lutheran, Presbyterian, Pentecostal, Episcopalian, Reformed, Church of Christ, etc.)
2 Roman Catholic
3 Mormon (Church of Jesus Christ of Latter-day Saints or LDS)
4 Orthodox (such as Greek, Russian, or some other Orthodox church)
5 Jewish
6 Muslim
7 Buddhist
8 Hindu
9 Atheist
10 Agnostic
11 Something else; Specify:_____
12 Nothing in particular

ASK IF SOMETHING ELSE OR NO RESPONSE TO RELIG (RELIG=11 or MISSING):
CHR Do you think of yourself as a Christian or not?

1 Yes
2 No

ASK IF CHRISTIAN (RELIG =1-4 OR CHR=1):
BORN Would you describe yourself as a born-again or evangelical Christian, or not?

1 Yes, born-again or evangelical Christian
2 No, not born-again or evangelical Christian

ASK ALL:
ATTEND Aside from weddings and funerals, how often do you attend religious services?

1 More than once a week
2 Once a week
3 Once or twice a month
4 A few times a year
5 Seldom
6 Never

ASK ALL:
RELIMP How important is religion in your life?

1 Very important
2 Somewhat important
3 Not too important
4 Not at all important

ASK ALL:

PRAY People practice their religion in different ways. Outside of attending religious services, how often do you pray?

1 Several times a day
2 Once a day
3 A few times a week
4 Once a week
5 A few times a month
6 Seldom
7 Never

ASK ALL:

FDSTMP_CPS Did anyone in your household get food stamps or use a food stamp benefit card at any time during 2015? *Do not include WIC benefits.*

1 Yes
2 No

ASK ALL:

WRKSTAT_GSS Last week, were you working full time, part time, going to school, keeping house, or what?

1 Working full time
2 Working part time
3 With a job, but not at work because of temporary illness, vacation, strike
4 Unemployed, laid off, looking for work
5 Retired
6 In school
7 Keeping house

ASK ALL:

REGISTERED Are you registered to vote?

1 Yes
2 No

ASK ALL:

PVOTE12A In the 2012 presidential election between Barack Obama and Mitt Romney, did things come up that kept you from voting, or did you happen to vote?

1 Voted
2 Did not vote (includes too young to vote)

ASK IF VOTED IN PVOTE12A (PVOTE12A=1):

PVOTE12B Did you vote for Obama, Romney or someone else?

1 Obama
2 Romney
3 Other candidate

ASK ALL:

VOTE14 In the 2014 midterm election, did things come up that kept you from voting, or did you happen to vote?

1 Voted
2 Did not vote (includes too young to vote)

ASK ALL:

FAMINC_CPS Which category represents the total combined income of all members of your FAMILY during the past 12 months?

This includes money from jobs, net income from business, farm or rent, pensions, dividends, interest, social security payments and any other money income received by members of your family who are 15 years of age or older?

1 Less than $5,000
2 $5,000 to $7,499
3 $7,500 to $9,999
4 $10,000 to $12,499
5 $12,500 to $14,999
6 $15,000 to $19,999
7 $20,000 to $24,999
8 $25,000 to $29,999
9 $30,000 to $34,999
10 $35,000 to $39,999
11 $40,000 to $49,999
12 $50,000 to $59,999
13 $60,000 to $74,999
14 $75,000 to $99,999
15 $100,000 to $149,999
16 $150,000 to more

ASK ALL:

ZIPCODE What is your zip code?

[ENTER NUMBER FROM 00000 to 99999]