Evaluating the Accuracy of Google Surveys

Katrina Sostek
Google Inc.
January 2019

Overview

Google Surveys is a market research platform that surveys internet and smartphone users. Our methodology whitepaper[1] explains how Google Surveys works and discusses its ability to mitigate different kinds of biases. This paper evaluates the accuracy of Google Surveys by comparing its survey results against benchmarks and other online survey platforms.

Since Google Surveys launched in 2012, several independent studies have been published on the accuracy of its results.[2] In addition, several large-scale studies[3] have been conducted on the accuracy and biases of non-probability online surveys by The Advertising Research Foundation (ARF) in 2014[4] and the Pew Research Center in 2016[5] and 2018[6]. Those 3 studies evaluated the results of multiple online survey platforms against benchmarks; The ARF evaluated 17 platforms, and Pew evaluated 9 in 2016 and 3 in 2018. None of these studies included Google Surveys due to its differences from other platforms, most notably Google Surveys' 10-question limit per survey.

In 2018, we replicated the 2018 Pew study as much as possible on the Google Surveys Publisher Network, which is a collection of 1,500+ sites, ~75% of which are News sites.[7] The most significant methodological difference was the length of the questionnaire; Pew's was ~60 questions, while ours was 6 separate surveys with 3-5 questions each. We grouped the questions into surveys by the 6 topics in the Pew study: Civic Engagement, Family, Financial, Political, Personal, and Technology. Of the 24 benchmark questions[8], we excluded 3 that were not in line with our policies,[9] leaving 21 questions. See Appendix C for a comprehensive list of all the ways in which our methodology differed from the Pew study.

The key findings from our experiments were

[1] g.co/SurveysWhitepaper (2018)
[2] A Comparison of Results from Surveys by the Pew Research Center and Google Consumer Surveys (2012)
How Representative are Google Consumer Surveys? (2013)
Survey Experiments with Google Consumer Surveys: Promise and Pitfalls for Academic Research in Social Science, Q&A (2017)
[3] A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples (2014)
[4] "Foundations of Quality 2.0: Launching What We've Learned," restricted content on thearf.org (2014)
[5] Evaluating Online Nonprobability Surveys (2016)
[6] For Weighting Online Opt-In Samples, What Matters Most? (2018)
[7] For more information on the Publisher Network, see g.co/SurveysWhitepaper.
[8] See the 24 benchmark questions and data sources here.
[9] The questions we excluded were US citizenship, gun ownership, and food allergies.

1. Google Surveys (GS) and Pew's overall errors versus benchmarks were similar, although Pew was able to reduce errors more effectively through weighting than GS — a 24% versus 10% reduction from the best-performing experimental weighting schemes.
2. GS and Pew both overrepresented people who are civically and politically engaged, a phenomenon that also occurs in telephone polls to a lesser extent.[10] GS' and Pew's samples differed in a few notable ways: GS underrepresented technology enthusiasts, while Pew overrepresented technology enthusiasts. GS' respondents were wealthier than benchmarks, while Pew's respondents was less wealthy than benchmarks.
3. As in the Pew study, we also found no "silver bullet" from adding more weighting variables. All the weighting variables we added — education, race/ethnicity, and voter registration — reduced errors for some topics but increased errors for other topics.

Topline Results

Pew's study examined the results of 3 online opt-in panel vendors in 2016 against different methodological combinations: 7 weighting methods, 2 sets of weighting variables, and 8 sample sizes from 2k to 8k in increments of 500. Our study examines the results from the GS Publisher Network in August 2018 using the raking weighting method, a set of 3 weighting variables, and sample size of 3k; see Appendix A for more details on the study design.

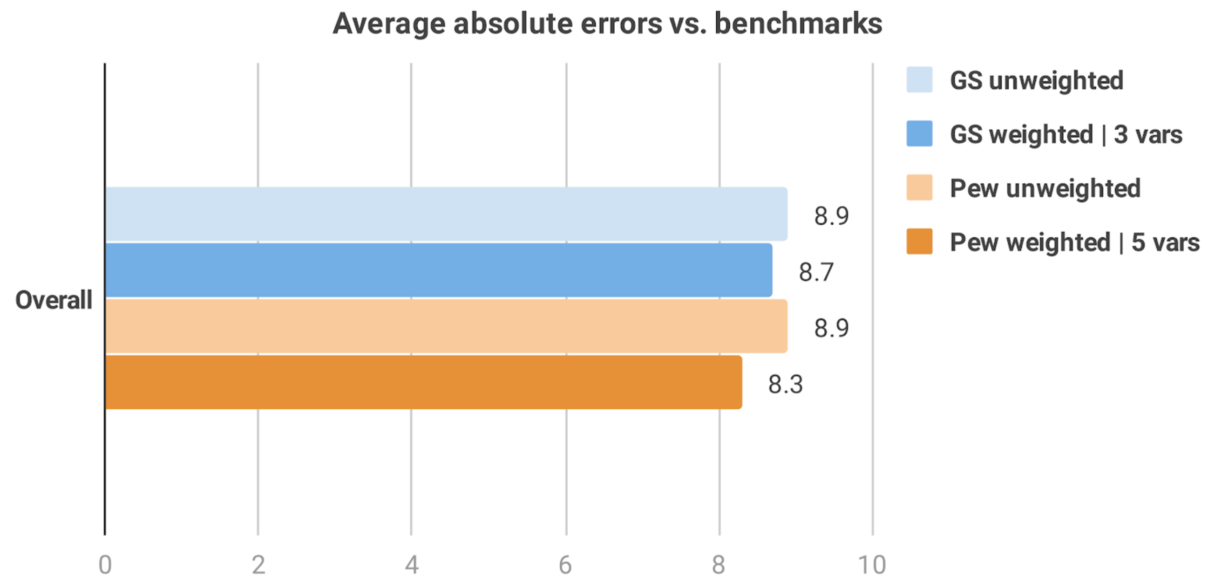
In the results below, we compare GS to only one of Pew's methodological combinations:

Study	Weighting method	Weighting variables	Sample size
GS	Raking	Age, gender, region (3 vars)	3k
Pew	Raking	Age, gender, region, race/ethnicity, education (5 vars)	8k

To simplify our comparison, we chose the raking weighting method because GS uses raking by default[11] and Pew found that raking performed comparably to other, more complex methods. We chose GS' 3 default weighting variables and Pew's most similar set of 5 instead of 9 weighting variables. Later in this paper, we'll compare our results to Pew's results using 9 weighting variables, which performed better than 5 variables. We chose a smaller sample size — 3k instead of 8k — because Pew found that larger sample sizes had diminishing returns; their 2k-response samples had essentially the same results as their 8k-response samples.

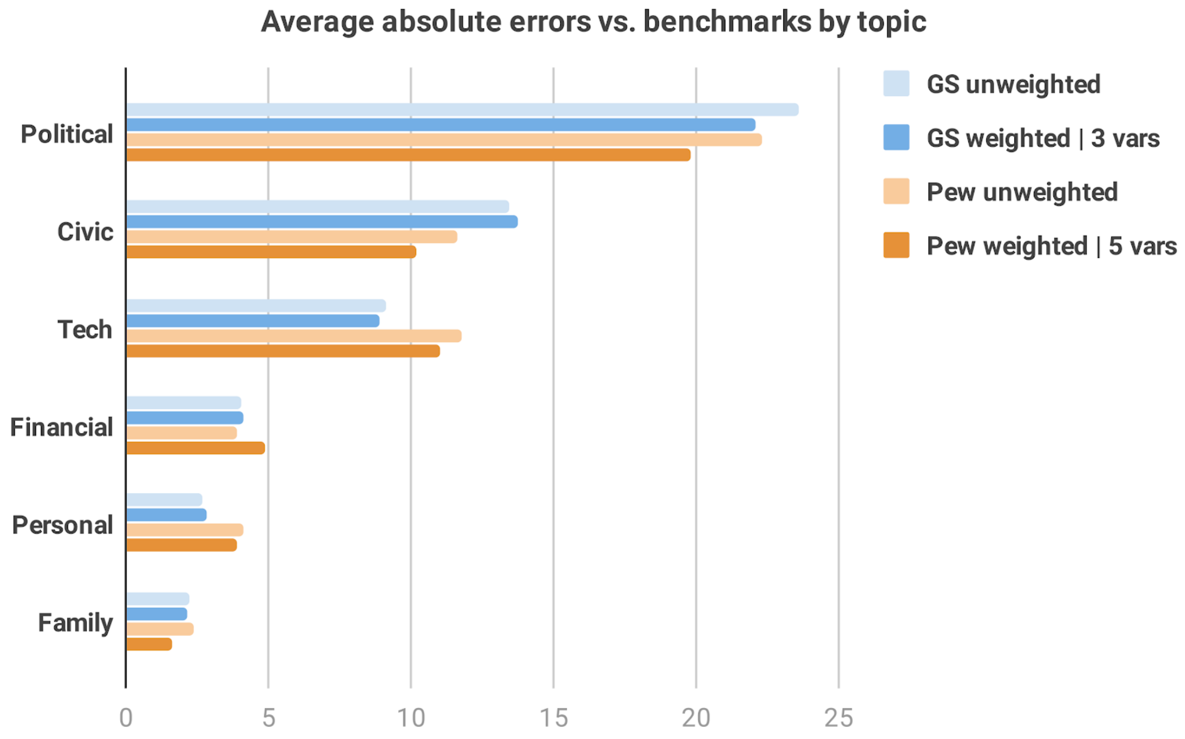
We calculated the average absolute errors vs. benchmarks in the same way as Pew at the overall, topic, and question levels.[12] Below, we compare overall errors between GS and Pew.

[10] What Low Response Rates Mean For Telephone Surveys (2017)
[11] For more details on GS' weighting methodology, see g.co/SurveysWhitepaper.
[12] For each question in a survey, we took the absolute value of the difference between each answer and benchmark value. To get the question-level errors, we took the average of the answer-level errors within each question. To get the topic-level errors, we took the average of the question-level errors within each topic. For the overall errors, we took the average of all the question-level errors.


Average absolute errors vs. benchmarks

Overall, GS' results were similar to Pew's when we compare weighting by 3 variables to 5 variables. The unweighted absolute errors were the same, and the weighted absolute errors differed by 0.4 percentage points — 8.3 versus 8.7. As a percentage of the unweighted error, Pew's weighting reduced errors by 7%, while GS' weighting only reduced errors by 2%.

Looking at the errors by topic, we can see more similar trends between GS and Pew.

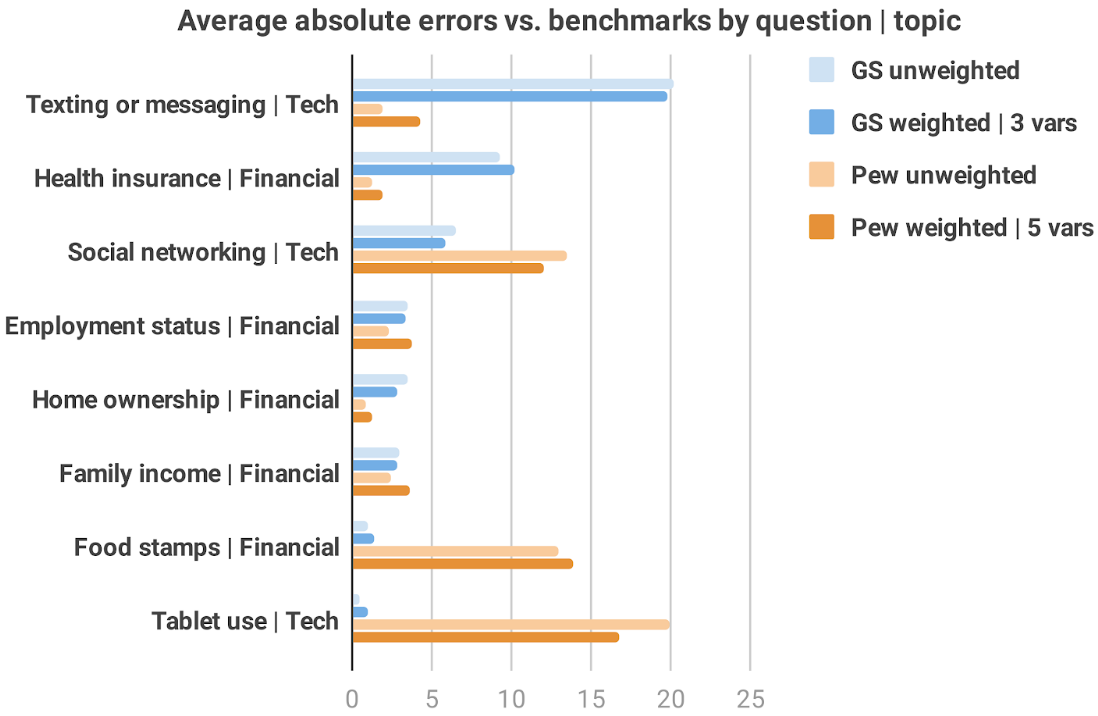

Average absolute errors vs. benchmarks by topic

GS' results were similar to Pew's results in that the Political, Civic, and Tech topics had the largest errors. Note that telephone surveys also have high errors for Political and Civic topics, although to a lesser extent; previous research has shown that non-response bias causes telephone surveys to overrepresent politically and civically engaged respondents.[13]

Weighting Pew's results reduced errors for all topics except Financial. Weighting GS' results reduced errors for half of the topics while increasing errors for Civic, Financial, and Personal.

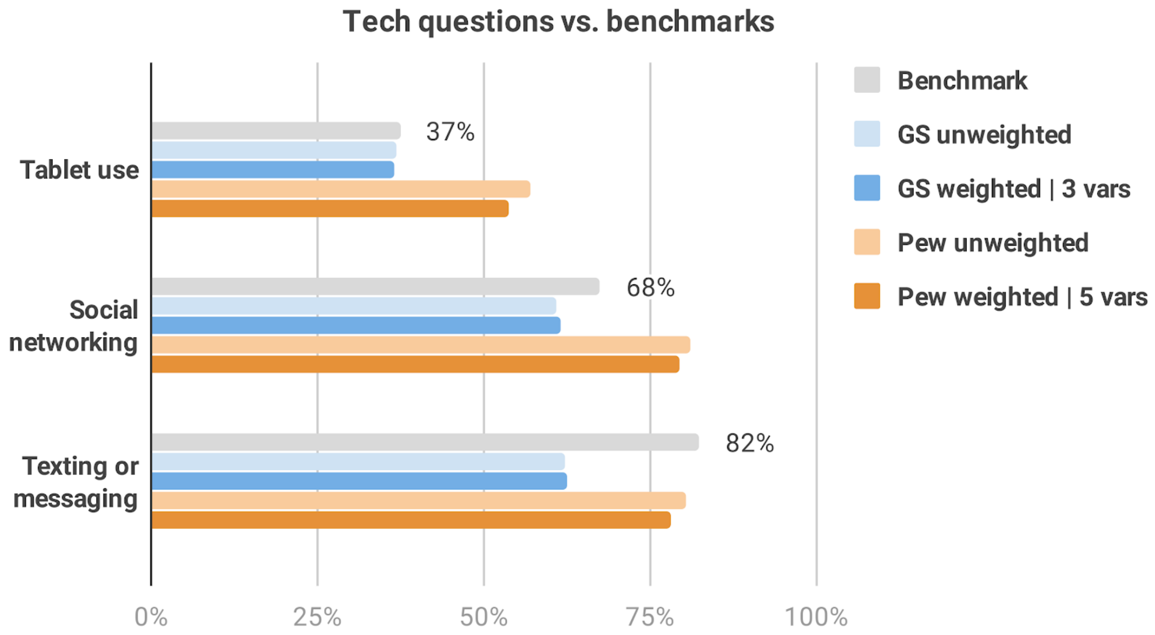
Benchmark differences

GS and Pew skewed the same way for the questions in the Civic and Political topics by overrepresenting respondents who were civically and politically engaged. In the 2 topics with the next highest errors, Tech and Finance, GS and Pew differed in how they skewed from benchmarks. Below, we examine the question-level errors within those topics.



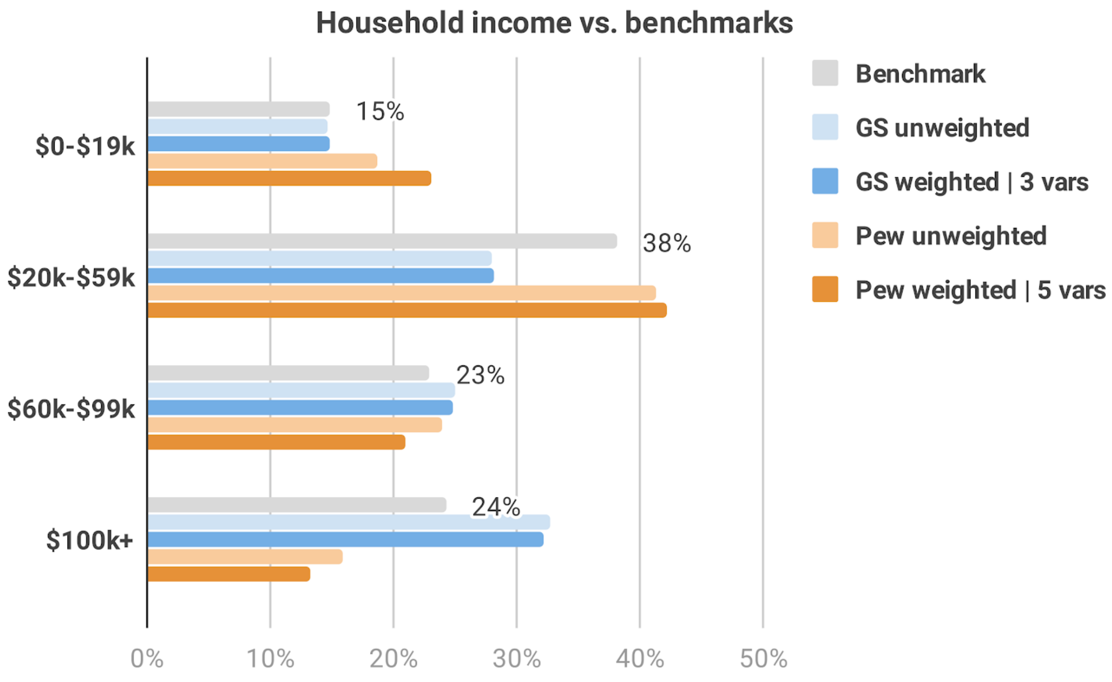
We can see that GS had the highest errors for texting/messaging, while Pew had the highest errors for tablet use. To fully understand the differences in these errors, we need to examine the answer-level results. We can first look at the results for Tech questions, specifically respondents who answered "yes" to using each of the following technologies.

[13] Assessing the Representativeness of Public Opinion Surveys (2012), What Low Response Rates Mean For Telephone Surveys (2017)



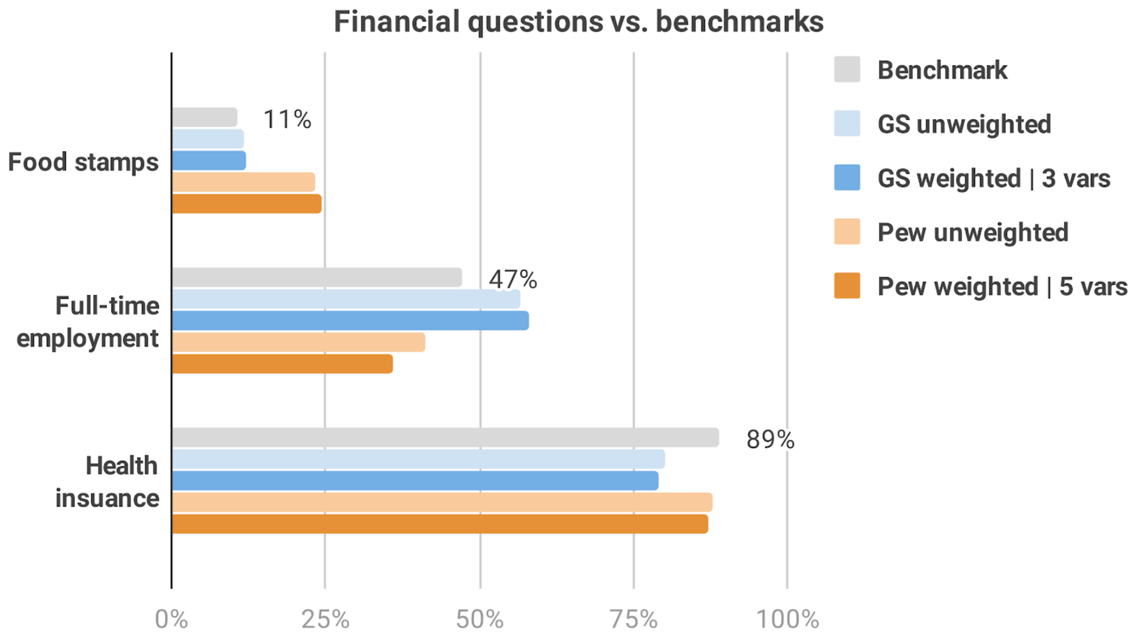
GS was close to the benchmark for tablet use, while Pew was close to the benchmark for texting/messaging. GS underrepresented respondents who use social networks and texting/messaging, while Pew overrepresented respondents who use tablets and social networks. These results show that not all online survey platform respondents are the same in terms of tech enthusiasm or adoption, which most likely stems from different recruitment methods. GS' respondents, despite answering surveys online, were not tech enthusiasts.

Next, we can see differences for the income question in the Financial topic.



Overall, GS's respondents were more wealthy than benchmarks, while Pew's respondents were less wealthy than benchmarks. GS underrepresented respondents with $20-$59k incomes and overrepresented those with $60k-$100k+ incomes. Pew overrepresented respondents with $0-$59k incomes and underrepresented those with $100k+ incomes.[14]

These income differences help explain the differences in a few other Financial questions.

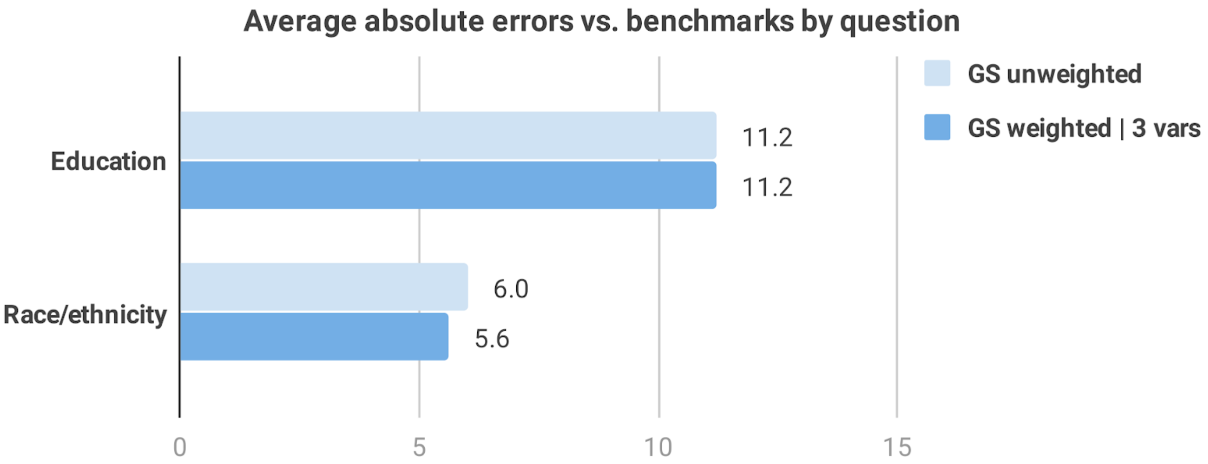


[14] Note that we combined the original 7 income answer choices into 4 to make the results easier to read.

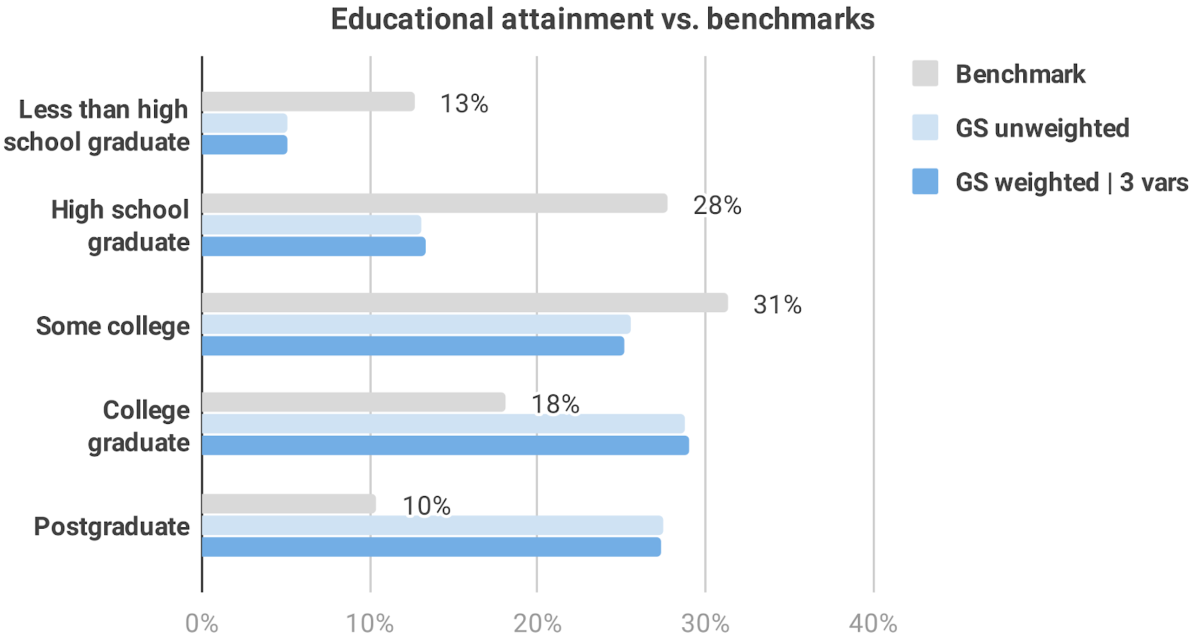
GS overrepresented people with full-time employment, which is consistent with GS' overrepresentation of wealthier people. In contrast, Pew underrepresented people with full-time employment and overrepresented people receiving food stamps. Interestingly, GS underrepresented people with health insurance, which seems at odds with the other results.

Sample composition

To better understand the demographics of GS' sample, we also ran a separate survey asking additional questions about educational attainment and race/ethnicity.

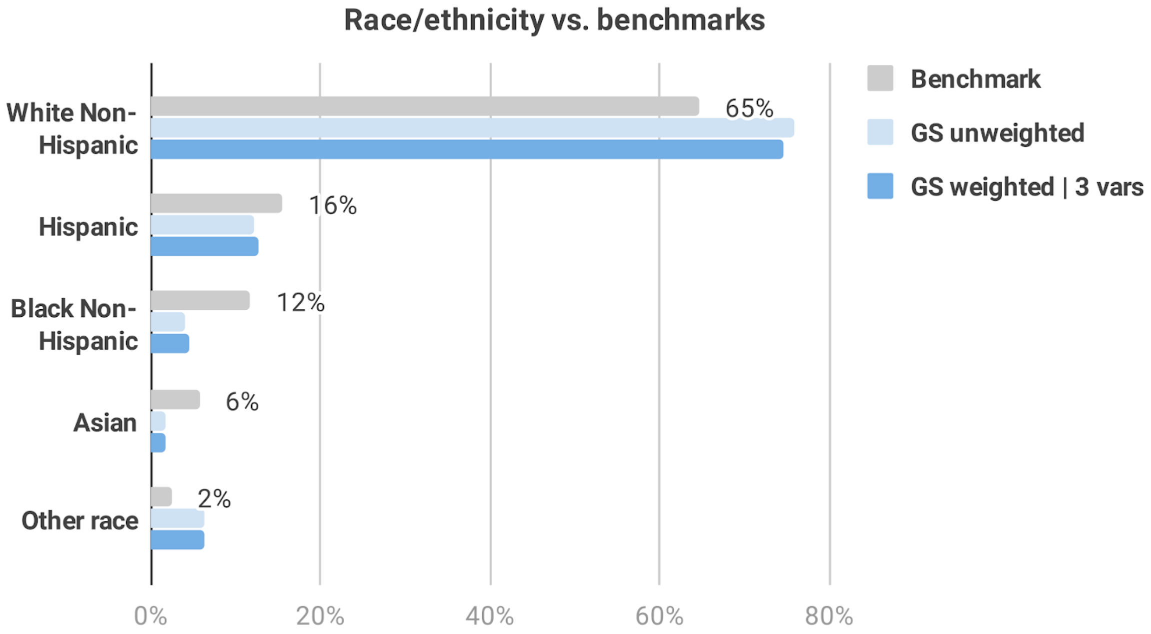


The source of the Educational attainment errors is clear from the answer-level results.



GS underrepresented people with less than a college degree by 5-15 percentage points and overrepresented people with college and postgraduate degrees by 11-17 percentage points.

We can also examine the answer-level errors for race/ethnicity.



GS overrepresented White Non-Hispanic people by 10 percentage points and Other race by 4 percentage points.[15] GS underrepresented Hispanic people by 3 percentage points, Black Non-Hispanic people by 8 percentage points, and Asian people by 4 percentage points.

Weighting experiment

The 2018 Pew study compared combinations of different weighting methods and weighting variables. Their paper concluded that raking performed comparably to more complex weighting methods, but the weighting variables that were chosen was more important than the method that was chosen. Their most accurate results came from weighting by 9 variables – 4 political variables in addition to the standard 5 demographic variables – which helped reduce the large errors in the Political topic. However, the Pew paper concluded, "even the most effective adjustment strategy was only able to remove about 30% of the original bias."[16]

[15] Note that Other Race included respondents who selected more than one race or entered an open-ended answer, which could have been used as a way to opt out of answering. Classifying multiple races as "Other Race" follows the way that the American Community Survey and Pew calculated their race/ethnicity results.

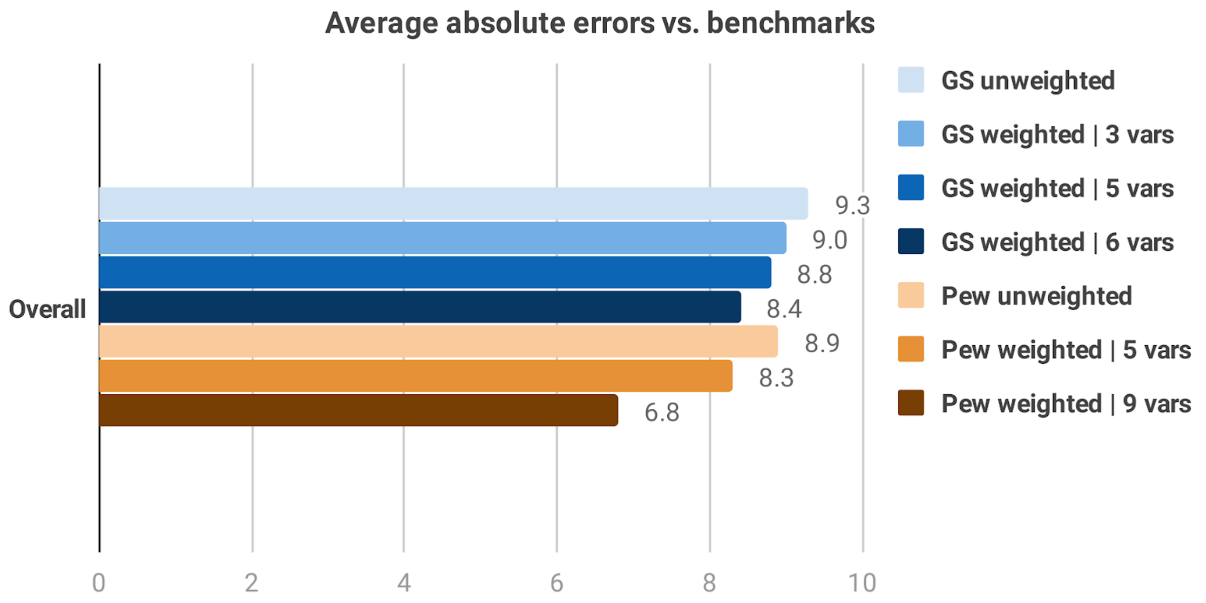
[16] Our results show that Pew removed only 24% of bias, which is due to our removing 3 benchmark questions.

By default, GS uses raking with 3 weighting variables: age, gender, and region. We ran an experiment in July 2018 to see if we could reduce errors by adding more weighting variables: education, race/ethnicity, and voter registration, which was the political variable that most reduced errors in the Pew study. We ran the same surveys as above with 4 additional questions at the end: Education, race/ethnicity (2 questions), and voter registration.

Study	Weighting method	Weighting variables	Sample size
GS	Raking	Age, gender, region (3 vars) + race/ethnicity, education (5 vars) + voter registration (6 vars)	3k
Pew	Raking	Age, gender, region, race/ethnicity, education (5 vars) + voter registration, political party identification, political ideology, identification as an evangelical Christian (9 vars)	8k

This experiment has a few limitations compared to the experiments Pew ran. We added more weighting variables by adding questions to the end of the survey, but we were limited by our inability to change the initial sampling for the survey, which still matched the distribution for age x gender x region. We were also limited in how many additional weighting variables we could add because of the 10-question limit; to get 9 weighting variables total, we'd need to add 6 more questions, but doing that would cause surveys to have more than 10 questions.

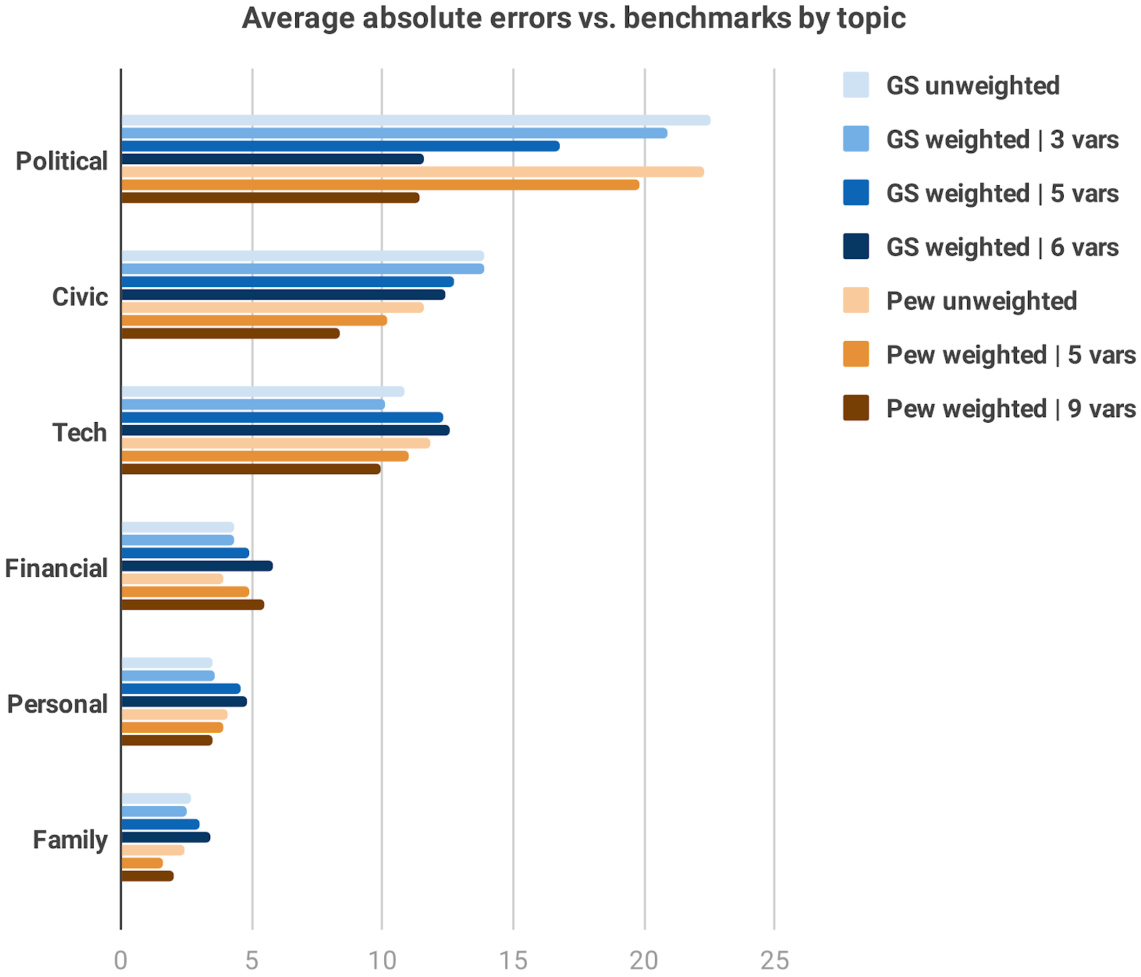
We can compare the results for GS and Pew using different sets of weighting variables.



Compared to GS' unweighted results, weighting by 3 variables achieved a 3% reduction in errors, while weighting by 5 variables reduced errors by 5%. Weighting by 6 variables — age, gender, region, race, education, and voter registration — reduced GS' errors by 10%.

Pew's weighting by 5 variables reduced errors by 7% compared to their unweighted results; only slightly more than GS' reduction of 5%. However, Pew achieved a much larger reduction of 24% from weighting using 9 variables vs. GS' 10% reduction from using 6 variables.

For both GS and Pew, the best-performing weighting schemes were effective because they reduced errors in the Political and Civic topics, which had the highest absolute errors.



Weighting GS by 5 and 6 variables reduced errors for the Political and Civic topics, but also increased errors for all 4 other topics. In comparison, weighting Pew by 5 and 9 variables only increased errors in one topic: Financial. As we can see, different weighting variables reduced or increased errors depending on the survey topic. Indeed, the Pew paper recommends "A careful consideration of the factors that differentiate the sample from the population and their association with the survey topic" when considering which weighting variables to use.

Another factor to consider when adding more weighting variables is how much variance the weighting scheme introduces. We can calculate the design effect[17] to assess how much variance has been introduced by weighting. If the design effect goes above 2, it's considered a warning that the weighting method is no longer worth the tradeoff between bias and variance. For GS, the design effect for 3 variables was 1.0, for 5 variables was 1.8, and for 6 variables was 2.1. So, the most effective weighting scheme was just barely over the limit for potentially introducing too much variance for the sake of reducing errors by 10%.

Appendices

Appendix A: Study design

Two rounds of surveys were run on the following dates:

1. Standard set of questions, August 1-3, 2018: Results
2. Additional questions for weighting variables added, July 25-26, 2018: Results

Each round of surveys was conducted as follows:

- 6-7 surveys were run: one survey per topic plus an additional demographics survey.
- 3 waves of each survey were run at the same time.
- 1.5k complete responses were targeted for each survey.
- We dropped responses with unknown inferred age, gender, or region because they cannot be weighted, which lowered the average number of completes for each survey to 1.1k. We did this for both unweighted and weighted results for consistency.
- For the analysis above, the three surveys were combined into a 3k-response sample.

GS guarantees no duplicate respondents within surveys but not between different surveys. When constructing the 3k-response samples, we removed duplicates. Within a single topic, the average percentage of respondents who got multiple surveys was 0.07%. Across all topics, the average percentage of respondents who got multiple surveys was also 0.07%.

In our analyses, we included partial responses when possible; i.e., results from respondents who didn't complete the entire survey. For surveys with no additional questions for weighting variables, the average dropoff rate from the first to last question was 8%. Adding education, Hispanic origin, and voter registration questions increased the dropoff rate to 21%.

[17] Using the calculation for design effect from Effects of Sample Design on Statistical Inference

The combined response-level results are included below with links to the individual surveys.

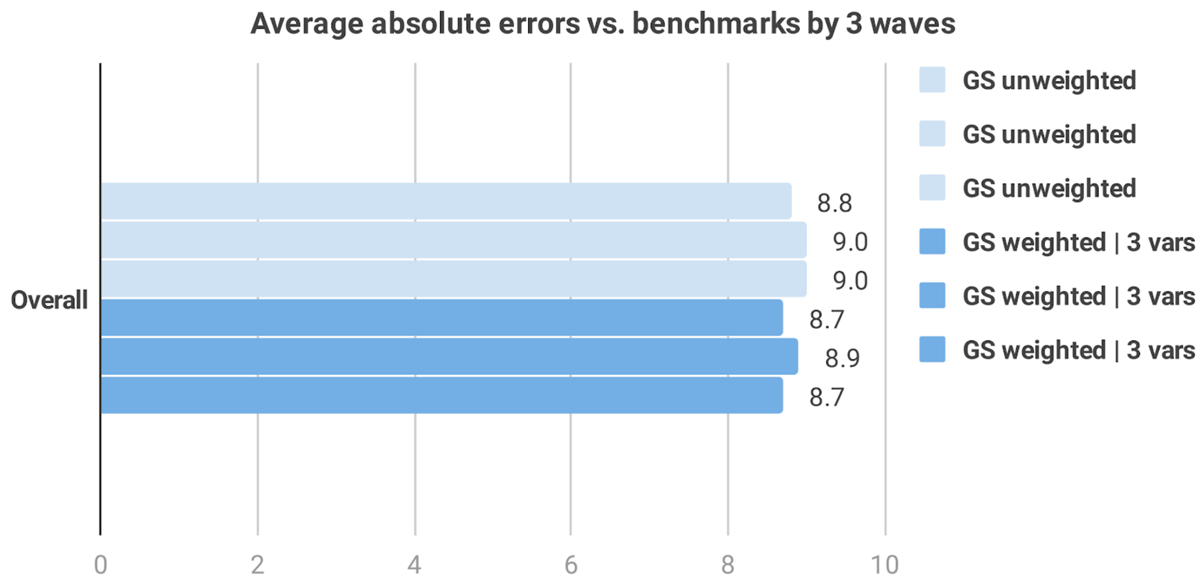
Survey topic	**August 1-3: Results, Surveys**	**July 25-26: Results, Surveys**
Civic	Combined, Waves 1, 2, 3	Combined, Waves 1, 2, 3
Family	Combined, Waves 1, 2, 3	Combined, Waves 1, 2, 3
Financial	Combined, Waves 1, 2, 3	Combined, Waves 1, 2, 3
Personal	Combined, Waves 1, 2, 3	Combined, Waves 1, 2, 3
Political	Combined, Waves 1, 2, 3	Combined, Waves 1, 2, 3
Technology	Combined, Waves 1, 2, 3	Combined, Waves 1, 2, 3
Demographics	Combined, Waves 1, 2, 3	

The per-question sample sizes in the combined results ranged from 3.3k to 4.8k. The smallest sample size was from the last question in a survey (completes). The largest sample size was from the first question (partials) in a survey with a high dropoff rate. The number of completes ranged from 3.3k to 3.5k, which we abbreviated to 3k in this paper.

We calculated weights for each question using the partial results available for that question, which GS does by default. The initial weights for each survey were calculated based on the first question. Weights for other questions were calculated by taking the first-question weights, filtering them to those who answered the other question, and renormalizing them so that the sum of the weights equaled the total number of responses to the other question. For more information about how weighting works in GS, please see g.co/SurveysWhitepaper.

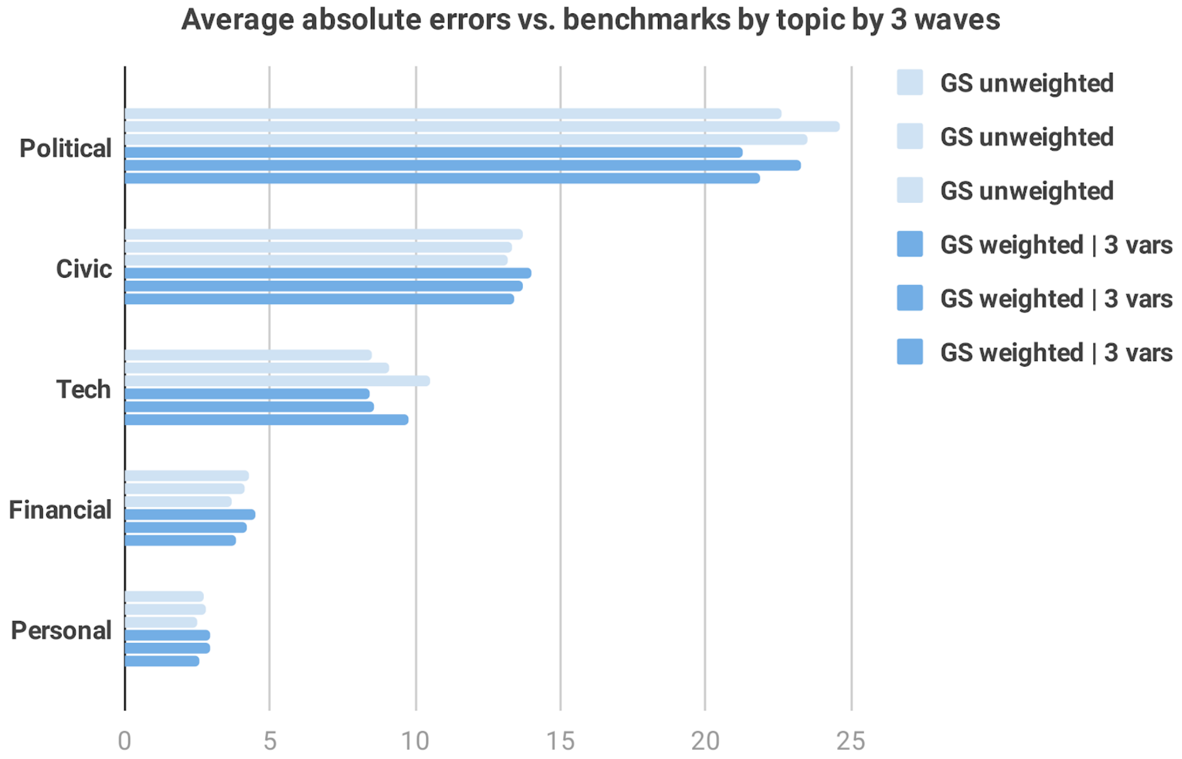
Appendix B: Consistency of results

Running multiple surveys at once allowed us to examine the consistency of their results. In the first weighting experiment, the 3 waves of 1.1k-response samples were highly consistent with each other for both the unweighted and weighted overall absolute errors.



For the 3 waves of surveys shown above, the overall errors for unweighted results were within a range of 0.23 and the weighted results were within a range of 0.21. The standard deviation for unweighted results was 0.12 and for weighted results was 0.10.

We can see the same pattern of consistency at the topic level.



The Political and Tech topics had the largest ranges for unweighted results at 2.0. The Political topic had the largest range for weighted results at 2.0. The standard deviation within topics was 0.1-1.0 for unweighted and 0.2-1.0 for weighted results; the Personal topic had the smallest standard deviation and Political and Tech had the largest standard deviation.

Appendix C: Survey design

The questions appeared in the following order in the surveys, following Pew's questionnaire:

- Civic: Talk to neighbors, trust neighbors, community group, volunteer (2 questions)
- Family: Marital status, house size, number of children
- Financial: Home ownership, health insurance, food stamps, employment, income
- Personal: Home tenure, military duty, smoking frequency (2 questions)
- Political: Contacted public official, voted 2012, voted 2014
- Technology: Tablet use, texting or messaging, social networking
- Demographics: Education, Hispanic origin, Race, Political party affiliation
- Example surveys: Civic, Family, Financial, Personal, Political, Technology, Demographics

Questions for additional weighting variables were added at the end in the following order:

- Education, Hispanic origin, Race, Voter registration
- Example surveys: Civic, Family, Financial, Personal, Political, Technology

In some cases, we changed the question and answer text due to

- Character limits for questions and answers
- Limits on the number of answer options
- Limits on the question types available; e.g., open-numeric questions

Summary of question and answer changes vs. Pew's questionnaire:

1. Community group (Civic): Removed reference to July 2015
2. Volunteer (Civic): Shortened question
3. Volunteer part 2 (Civic): Removed reference to June of last year, shortened question
4. Household size (Family): Changed open-ended numeric box to "1", "2", "3", "4 or more"
5. Children (Family): Changed open-ended numeric box to "0", "1", "2", "3", "4 or more"
6. Home ownership (Financial): Shortened the answers "Owned by household member with mortgage/loan" and "Owned by household member, no mortgage/loan"
7. Employment status (Financial): Shortened "With a job, but not at work (sick, vacation)"
8. Income (Financial): Coarsened the number of answer options from 16 to 7
9. House tenure (Personal): Shortened answers, "No, outside the United States/Puerto Rico" and "No, but in the United States/Puerto Rico"
10. Military duty (Personal): Shortened "Only for training in Reserves/National Guard"
11. Education (Demographics): Coarsened the number of answer options from 14 to 6

12. Race (Demographics): Changed “Some other race” to an open-ended text box, following the American Community Survey
13. The following questions showed their answers in a fixed order: Talk to neighbors, Trust neighbors, Income, Household size, Children, Smoking frequency part 2, Income, Age, Education, Hispanic origin, Race, Political party affiliation
14. The rest of the questions randomized the order of their answer options

Sampling and weighting:

- Sampled by the joint distribution of age, gender, region; no race/ethnicity or education
- Raking by the marginal distributions only, no pairs of variables (e.g., age and gender)
- The age, gender, and region population distribution used for sampling and weighting was from the 2015 Current Population Survey’s Computer and Internet Use Supplement

Language:

- GS ran all surveys in English
- Pew’s surveys were available in both English and Spanish

GS’ surveys ran 2 years after Pew’s surveys ran in June-July 2016, which may have affected results such as how well respondents remember if they voted in 2012 or 2014.

We compared our results to the same benchmarks as Pew to reduce the differences in our methodology; note that Pew’s benchmark data sources are from 2013-2016.

The microdata from Pew’s 2018 study is available for download at http://www.pewresearch.org/methods/dataset/2016-online-opt-in-comparison-study. Estimates that were not originally published in Pew’s 2018 report were provided by request.

Acknowledgements

Thanks to the following colleagues at Google for their insights and expertise: Carolin Jahn, Michael Brandell, Shaun Lysen, and Yongwei Yang.

In addition, many thanks to Art Owen and Andrew Mercer for their help along the way. Art Owen helped as a consultant for Google and not as part of his Stanford responsibilities.