# Historic Turnout Project: Methodological Appendix

To assess whether American democracy gives all citizens equal political voice, it is important to understand differences in voter turnout across demographic groups and how the composition of the electorate has changed over time. In this project, we have developed transparent, methodologically robust, temporally comparable estimates based on publicly available data that allow the public to understand how turnout varies between groups, over geography, and across years.

## Previous Approaches

Traditionally, there have been three widely used sources of information on the demographic composition of the electorate.

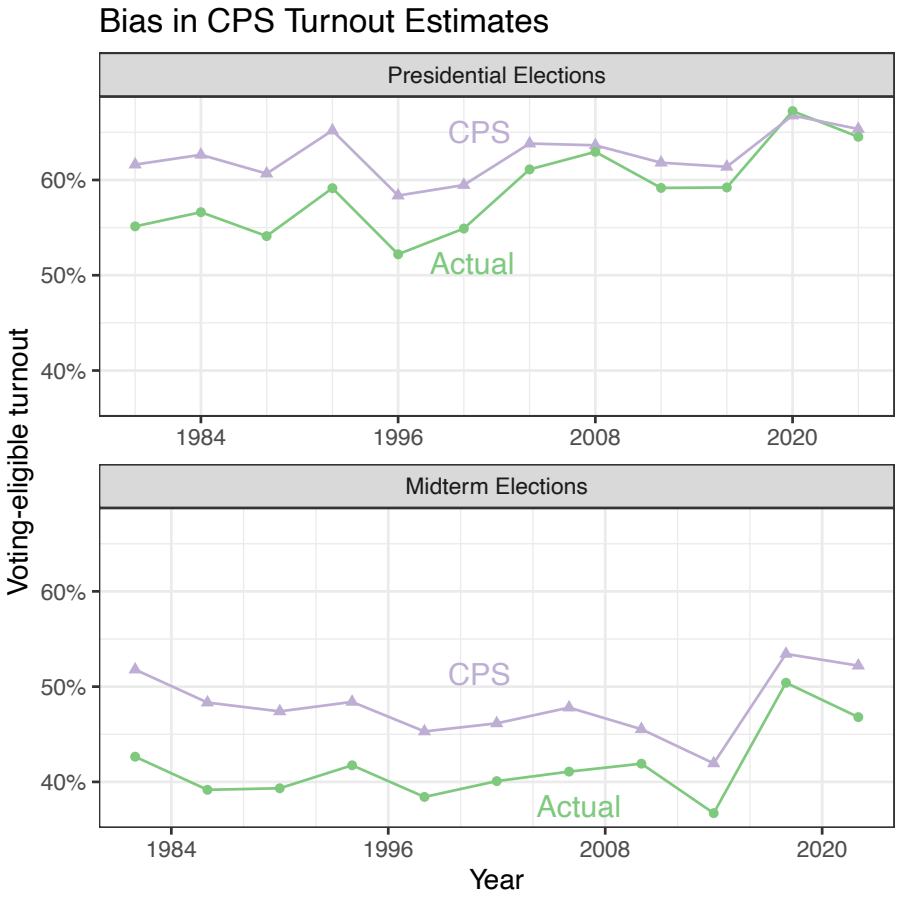
Current Population Survey

Perhaps the foremost source of information on voter turnout is the U.S. Census's Current Population Survey (CPS), which includes a November Supplement that asks respondents whether they voted in the last election. Because it is a based on a relatively large probability sample with high response rates, the CPS is often considered the "gold standard" for studying turnout, especially over time.[1] Because the CPS turnout estimate is based on the self-reported vote of those who choose to participate in the survey, however, it suffers from several well-known problems.

First, because the CPS relies on self-reported vote, it tends to overestimate actual turnout by at least a few percentage points. This is despite the fact that the CPS's official turnout estimates make the strong assumption that all respondents with missing self-reported vote are non-voters.[2] CPS estimates of turnout tend to be particularly high in midterm elections (see figure below).

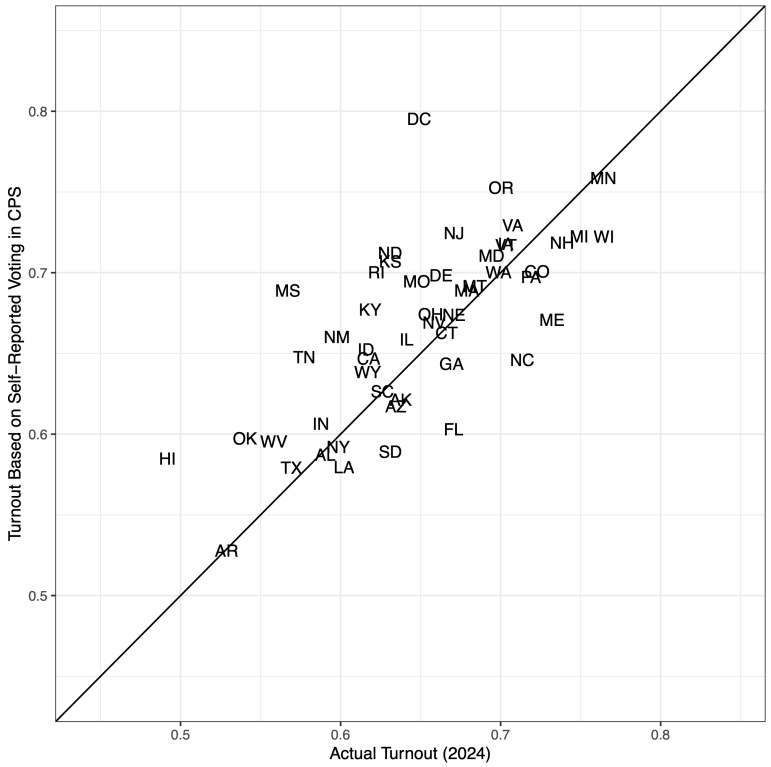
[1] Cohn, Nate. 2013. "The New Census Data That Should Terrify Republicans." *New Republic*, May. http://www.newrepublic.com/article/113160/november-2012-census-data-obamas-coalition-will-hold-together.

[2] Hur, Aram, and Christopher H. Achen. 2013. "Coding Voter Turnout Responses in the Current Population Survey." *Public Opinion Quarterly*, no. 4: 985–93.




In 2024, this tendency to over-estimate actual turnout was relatively muted at the national level, but it was very prominent in many states and especially the District of Columbia. The figure below plots the state-level turnout rates from the CPS against turnout rates based on actual vote counts in the 2024 presidential election.[3] While these two are certainly correlated, there are also substantial differences. For example, turnout in DC was almost 25% lower than you'd think based on the CPS. Such geographic biases in turn distort our understanding of turnout among demographic subgroups that may be concentrated in each area (e.g., Black Americans in DC).

[3] McDonald, Michael. 2025. "National Voting Eligible Population Turnout Rates, 1789-Present (V1.3)." https://election.lab.ufl.edu/dataset/national-vep-turnout-rates-1789-present-v1-1-3/. McDonald, Michael P. 2002. "The Turnout Rate Among Eligible Voters in the States, 1980–2000." *State Politics & Policy Quarterly* 2 (2): 199–212.




In addition, due to differential nonresponse correlated with education, White Americans without a college degree tend to be underrepresented in the CPS relative to the more authoritative American Community Survey (ACS).[4] In some states the discrepancy between the CPS and ACS can exceed 5 percentage points. This can further bias estimates of turnout at the subgroup level.

Voter Files

A second source of information is administrative voter files. These have been widely used in recent years. For instance, Catalist has released authoritative reports in recent years that examine how turnout and the party coalitions varies from election to election.[5]

Voter files do not include information about many demographic attributes, most notably race. These attributes are instead inferred from names and auxiliary sources of information using proprietary data and statistical models. What's more, high-quality voter files are unavailable prior to about 2008. As such, this data source cannot tell us how the electorate has changed over a longer time horizon.

---

[4] Cohn, Nate. 2013. "The New Census Data That Should Terrify Republicans." *New Republic*, May. http://www.newrepublic.com/article/113160/november-2012-census-data-obamas-coalition-will-hold-together.

[5] Catalist. (2024). What happened in 2024: An analysis of the 2024 presidential election. Retrieved September 12, 2025, from https://catalist.us/whathappened2024/

Ecological Inference based on Election Results

Third, we could use ecological inference based on actual election results.[6] This approach is widely used to estimate racial differences in turnout at the local level. However, this method **a)** requires fine-grained data on voting and demographics at the precinct-level that is not available over longer time periods, **b)** requires strong assumptions about the homogeneity of voting patterns within demographic groups across geography, and **c)** is not useful for making inferences about groups that don't vary across geographies (e.g., men and women).

## Our Methodology

With these kinds of issues in mind, we set out to build a methodology that could more accurately reflect the American electorate. There are several distinctive features of our approach:

- Based on publicly available data and has a fully transparent methodology
- Synthesizes multiple sources of data – leaning into the relative strengths of each to inform the others
- Able to generates estimates for major demographic subgroups at the state-level
- Comparable over elections – going from 1980 to the present
- Incorporating uncertainty into all the outputs in a way that reflects the uncertainty of data

Specifically, our estimates of turnout and the composition of the electorate take advantage of several sources of information, each with its own advantages and disadvantages:

- State Elections Offices provide data on the number of votes cast for each office, data which are aggregated to the county level by uselectionatlas.org and other sources.
- The American Community Survey is a multi-mode survey conducted by the Census Bureau based on a probability sample of 3.5 million households. The ACS enables estimates of the size and demographic composition of the population, but it does not ask respondents whether they voted. However, its population estimates can be combined with administrative data on votes cast to produce estimates of voter turnout at the state and substate level.
- The Current Population Survey is an in-person survey conducted by the Census Bureau based on a multistage probability sample of approximately 60,000 households (not including the institutionalized population). The CPS November

[6] King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.

Supplement asks respondents whether they voted in the most recent election. The CPS is much smaller than the ACS, but it still yields relatively large and representative samples of each state (though educated respondents are modestly overrepresented relative to the ACS). These samples can be used to estimate turnout by demographic category as well as by geography unit. The CPS' main disadvantage is that, due to respondents' tendency to overreport having voting, it generally overestimates aggregate voter turnout relative to estimates based on administrative and ACS data.

- The Cooperative (Congressional) Election Study is a national stratified survey with an opt-in online sample of about 50,000 respondents that has been fielded each election cycle since 2006. The advantage of the CES is that, in addition to asking respondents whether they voted, self-reported turnout is validated against official state voter files. The main disadvantage is that the CES sample overrepresents the politically engaged, resulting in large overestimates of voter turnout even for validated vote.

By incorporating all these sources of information, our procedures for estimating turnout mitigate the limitations of relying on any one source in isolation.
The estimation proceeds in several steps:

- **Multiple imputation:** For each year, the CPS and CES samples are concatenated. Validated vote is observed for most CES respondents but is entirely missing in the CPS sample. The missing values of validated vote are multiply imputed 20 times using the MICE algorithm.[7] A multilevel model with random effects by state is used for validated vote, with self-reported vote and demographic/geographic attributes as predictors. Other missing covariates are imputed with predictive mean matching. Imputation of validated vote largely addresses overreporting of turnout, the prevalence of which differs by demographic and geographic characteristics.
- **Calibration to state demographic benchmarks:** Each of the multiply imputed CPS datasets is raked to match ACS-derived targets for race, education, age, and gender in each state.[8] The raking ensures that the CPS is demographically representative of each state, most importantly with respect to education.
- **Calibration to substate turnout benchmarks:** Each of the multiply imputed CPS datasets is raked to match administrative and ACS turnout benchmarks at the most granular geographic level available in the CPS (either county or metropolitan statistical area). Two versions of the calibration are performed, by self-reported vote and, for surveys after 2006, imputed validated vote. The calibration ensures that estimates of either self-reported or validated turnout in each substate geography exactly match administrative benchmarks. Doing so at the substate

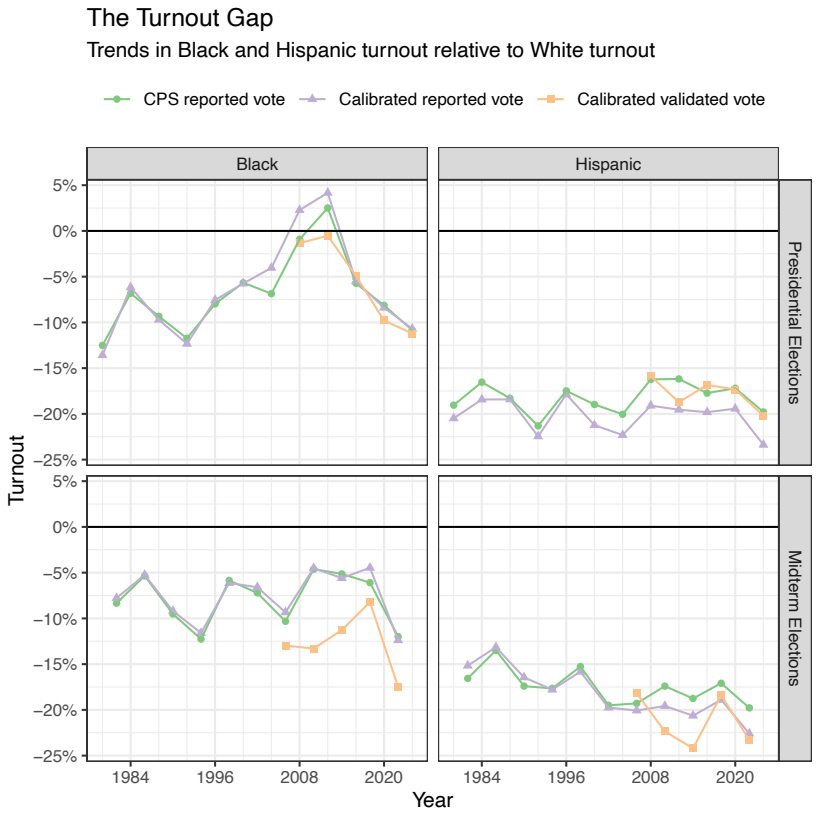
[7] van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

[8] Lumley, Thomas S. 2010. *Complex Surveys: A Guide to Analysis Using R*. Hoboken, NJ: Wiley.

rather than state level incorporates information contained in the correlation between turnout and the demographic composition of substate geographic units.[9]

In short, step (3) ensures that CPS turnout estimates match aggregate totals, step (2) ensures that the CPS is demographically representative, and step (1) accounts for differential overreporting across states and demographic groups. All analyses take into account the uncertainty in the imputations.[10]

The figure below compares the turnout gap across racial groups calculated using **a)** the reported CPS estimates, **b)** CPS estimates adjusted using the correction recommended by Hur and Achen (2013), and **c)** our final turnout estimates that incorporate all the information discussed above. In general, our final approach yields a larger turnout gap between White and Black Americans than simpler methods.




[9] Hur, Aram, and Christopher H. Achen. 2013. "Coding Voter Turnout Responses in the Current Population Survey." *Public Opinion Quarterly*, no. 4: 985–93.

[10] Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

## How Do We Know our Approach Works?

We have taken several approaches to validate our measures.

First, our multiply imputed estimates of validated turnout in the CPS from 2006-2024 are consistently closer to state-level turnout benchmarks than the raw CPS data reported by the Census.

Second, our calibrated estimates of the electorate match demographic benchmarks in each state. They also perfectly match administrative benchmarks of voter turnout at the substate level. As such, we know that our estimates are accurately capturing both the right number of people in each demographic group and the right numbers of voters.