American Datasets, Minority Research

Luke Bellinger

Race and American Politics Final Paper, Duke University

6/21/2025

1 Introduction

Many Race and Ethnic Politics scholars use datasets that are not oversampled and have deep biases when applied to minority populations. Most obviously, the CES does not interview in Spanish. Yet, various articles use this data to make inferences about Hispanic populations (Fraga et al. 2025; Derek Wakefield 2025). Why? Well, among other things, the CES has a larger n, and forms a complete time-series (Derek Wakefield 2025). Oversampled datasets are often one-off, or are relatively new, such as the CMPS, which has only had four waves to date. Despite use of American Politics datasets becoming more common in Race and Ethnic Politics work, articles rarely consider the underlying biases in sampling when using datasets not designed for Race and Ethnic Politics (Fraga et al. 2025), though exceptions exist in some online appendices (Tokeshi 2023; Corral and Leal 2024) and technical reports (Deshpande and Cha 2022). I examine biases in American Politics datasets de novo with a new framework of evaluations specific to race and ethnic politics work. This framework takes into account descriptive trends and regression models. I find that the CES, and to a lesser extent, the ANES, undersample Hispanics, especially in highly Hispanic states, and oversample Whites. I also find that language is not a clear driver of this disparity. Finally, I find that the CES and ANES underestimate the effect of race on outcome variables when included as an explanatory variable in a regression on a whole sample of the American population, but regression coefficients in subsamples of the CES and ANES are not significantly different from oversamples form the CMPS. Race and Ethnic Politics scholars should continue to use American Politics

datasets when no better options are available, but using American Politics datasets is acceptable for some types of regression analysis, though it is still important to know the existing biases in the American Politics data before using it.

2 Literature Review

At the front of the paper, it is important to recognize three things. First, the paucity of data for Race and Ethnic Politics is because of problems in funding. Funders don't want to fund Race and Ethnic Politics data. I'm aware that this paper criticizes Race and Ethnic Politics scholars making the best out of a bad data situation. Second, by the sole virtue of being a statistical minority, it is difficult and expensive to sample non-White groups. Third, both Hispanic and Asian-American categories are loosely formed social constructions, which is why I evaluate biases in subgroup ethnicities as well (Wong et al. 2011; Derek Wakefield 2025). This paper is intended to inform Race and Ethnic Politics researchers using American politics datasets, not necessarily to discount or discourage their utilization outright. Even though American politics datasets may be the best data available to answer many questions, it is worth knowing the biases in using data not designed for use in Race and Ethnic Politics work.

Use of American Politics datasets in recent Race and Ethnic Politics work

Various Race and Ethnic Politics scholars have used American Politics datasets for their work. The Cooperative Election Study (CES) (Schaffner et al. 2023), formerly the Cooperative Congressional Election Study, is most commonly used, owing to its frequency and large n-size and geographically stratified sample (Wilkinson et al. 2015; Fairdosi and Rogowski 2015; Stout and Garcia 2015; Chan 2022; Kellstedt and

Guth 2021; Benjamin and Carr 2022; Cuevas-Molina 2023; Watts 2024; Geiger and Reny 2024; Brown et al. 2025; Fraga et al. 2025). The American National Election Studies' time-series (American National Election Studies 2022) is used less frequently (Abrajano and Panagopoulos 2011; McKenzie and Rouse 2013). Both form complete time series, and have been used as parts of time-series analyses (Enders and Thornton 2022; Derek Wakefield 2025), rare in Race and Ethnic Politics. The Race and Ethnic Politics alternative is the Collaborative Multiracial Post-election Survey (CMPS) (Frasure et al. 2024), which contains oversamples for Black, Hispanic, Asian, and now American Indian (Foxworth et al. 2025) populations. It is commonly used in Race and Ethnic Politics work because of its oversamples (Masuoka et al. 2019; Irizarry 2024; Slaughter et al. 2024). The CES contains enough racial minority respondents to match the or exceed the CMPS for certain groups, (CES Black n = 6952, Hispanic n = 5180, Asian n = 1831; CMPS Black n = 4613, Hispanic n = 3873, Asian n = 3836), which may make it tempting to use the CES for minority subsamples, even if the CES methodology documents warn against it (Schaffner et al. 2025, 18), because small mistakes can become very large when analyzing a small subsample (Ansolabehere et al. 2015). Accordingly, some studies that use multiple datasets do obiter dicta analyses comparing the CES to oversampled datasets (Tokeshi 2023; Corral and Leal 2024). Others combine and reweigh oversampled datasets with American politics datasets (Derek Wakefield 2025). Though Race and Ethnic Politics scholars are generally aware of possible biases when using non-oversampled data, many do so anyway. Thus, a comprehensive analysis between American Politics and oversampled datasets is necessary to identify and illustrate the differences between

using American Politics and Race and Ethnic Politics data.

Given that American Politics data is not made for use in Race and Ethnic Politics, I center my literature review around discussion of two topics that I think are especially important when sampling racial minorities: language proficiency and geography.

2.1 Language

The CES completed two Hispanic oversamples in 2010 and 2016, in which Spanish-language questionnaires were translated. All other CES interviews were done solely in English (Personal Communication with Brian Schaffner). Having interviews in Spanish deeply matters for any analysis of Hispanic politics. An analysis of a biased exit poll from CNN/Edison in 2016 finds that when there are no Spanish interviewers and when precincts for exit polling are not selected carefully, estimates of Hispanic vote choice can be deeply biased, in the case of the Edison poll, by an average of 15% in favor of Trump (Barreto et al. 2017, 220). Exit poll results can be affected by languages which interviews are available in, selection of precincts in segregated cities, and post-survey weighting techniques (Barreto et al. 2006). Response rates and representativeness among Hispanics in hard-to-reach areas, such as Texas *colonias*, can be boosted by competent Spanish interviewers and probability samples generated from satellite imaging, where no addresses or post offices are available (O'Hegarty et al. 2010). Speaking Spanish is a strong predictor of demographic variables, such as religion, and there are deep differences in those demographics between surveys that interview in Spanish and those that do not (Perl et al. 2006). Weighting can be used

to correct some of this bias, but it is not a substitute for valid data collection (Perl et al. 2006). For Asian Americans, the divide in language proficiency is significant enough to be used as a measure of acculturation, and foreign-born Asians who speak English have significantly more access to healthcare and resemble the demographic characteristics of US-born Asians more closely than foreign-born Asians who do not speak English (Lee et al. 2011). When Chinese- and Vietnamese-language surveys are fielded in the United States, researchers are able to garner similar response rates as English-language surveys with no more missingness of data (Ngo-Metzger et al. 2004). Fielding multilingual surveys for Asian Americans, Native Hawaiians, and Pacific Islanders is challenging as translations are often only done in the most spoken languages, and smaller languages are often ignored (Lee et al. 2022). Similarly, translations are expensive and do not take into account literacy levels or colloquial speech in these communities (Lee et al. 2022).

Language also has a broader effect on politics and political behavior for groups with large non-English speaking populations. The English language is often seen as the dominant language in the United States, and many people see English as integral to national identity and support laws to restrict the rights of non-English speakers and to mandate the use of English in America (Schildkraut 2013). Voting Rights Act (VRA)-mandated Spanish-language ballots and election assistance in California not only increased overall turnout for primary elections in that state, but also changed electoral outcomes in local policy (Hopkins 2011). However, Spanish-language Get Out The Vote (GOTV) efforts were only found to boost mobilization in low vote-propensity voters and Spanish-speaking voters, whereas English-language appeals

had a broader effect in boosting voting among all groups (Abrajano and Panagopoulos 2011). When candidates use Chinese-language names on campaign signs, they are perceived more positively by Chinese-American voters, regardless of whether the candidate themself is Chinese or not (Yeung and Wamble 2024). English-speaking Hispanics are less likely to prefer pan-ethnic labels than Spanish speakers (Chong 2019, 34), whereas non-English-speaking Asians are more likely to prefer national origin candidates than English-speaking Asians (Chong 2019, 42). Consumption of Spanish-language media is correlated with liberal attitudes on immigration and increased group consciousness, and "Hispanics who rely on Spanish-language media have significantly different views about politics in the United States compared to those Hispanics who primarily use English-language media" (Kerevel 2011, 529). Even the structure of a language can have an effect on public opinion, those who speak non-gendered languages are more likely to support gender equality than people who speak gendered languages (Perez and Tavits 2022). People interviewed in a minority language are more likely to view politics through an ethnic lens and prioritize minority issues, and the salience of these issues is affected by the ability to speak said minority language (Perez and Tavits 2022, 108-109).

2.2 Geography

Geography is important to race and ethnic politics because of the history of segregation and redlining. The government and private agencies segregated neighborhoods by race, leading to differential outcomes for housing prices (Rothstein 2017). Home Owners' Loan Corporation (HOLC) redlined neighborhoods still experience differen-

tial outcomes in health outcomes, pollution, income, and many other facets of life (Swope et al. 2022). Many cities remain divided by segregation on both race and class lines, such as Tucson where Hispanic population distribution is autocorrelated (*Moran's I = 0.859**)) as is income (*Moran's I = 0.551**)), or Harrisburg, where Black population is autocorrelated (*Moran's I = 0.668**)) (Frank 2003, 153, 155). inbook go back for citation For Black voting districts, though race explains 60% of variation in district-level voting patterns, geography explains a remaining 30% of variation, and districts with high racial polarization are concentrated in the South (Kuriwaki et al. 2024). This means that racially polarized voting is deeply contingent on place and geographic context. Racially polarized voting is prevalent in urban areas and is robust to changes in election rules meant to mitigate this clustering (McDaniel 2018). There are also unique effects of the urban-rural divide on race and ethnic politics. For White Americans, rural resentment and rural consciousness push rural voters to vote in opposition to urban voters in an us versus them dynamic (Cramer 2016). Yet the same voting pattern is not seen among rural voters of color, whose vote choice closely matches that of urban voters of color (Brown et al. 2025). However, rural voters of color are slightly more likely to hold conservative views, especially on social issues (Brown et al. 2025). Race and ethnicity and state politics are closely linked; racial diversity accounts for much of the diversity in the political culture of a state (Hero and Tolbert 1996). Conversely, state politics can also have an effect on racial minorities' public opinions and views of themselves (Jiménez et al. 2021).

To correct for non-representativeness of samples, researchers often use multilevel regression and post-stratification (MRP) weights to weight respondents' leverage in

counts and models by their demographic characteristics as opposed to a benchmark known population distribution, usually from the census. MRP is made up of two steps: multilevel regression and post-stratification. Multilevel regression refers to the fitting of a multilevel logit or probit model against demographic and geographic variables against an outcome variable of interest (Park et al. 2006).

The difference between stratification and post-stratification is that in stratification, strata are created based on complete frame data before sampling and post-stratification is based on aggregated strata after sampling (Kolenikov 2016).

3 Theorizing Datasets

We are entering an era in which no one wants to fund any research related to race. One could argue that race was already discouraged as a subject of inquiry prior to this administration, but there has been a substantive sea change in funding patterns for social science research, especially in relation to race and ethnicity. Despite criticizing various scholars' work using better funded American politics datasets, I do want to leave this paper with actionable advice for scholars and practitioners for how to do inquiry in a changing political climate. Survey experiments are not vastly more expensive per respondent than observational data, but are far more powerful for causality at smaller n sizes (Laird and White 2021; Jiménez et al. 2021). Qualitative research can produce reasonably robust results (Brady and Collier 2010; King et al. 2021), for less money and very small n, and has been used to produce some extremely high-quality Race and Ethnic Politics research (Simpson 1998; Harris-Lacewell 2010). The big-n dataset is not the end-all-be-all of empirical research, or even quantitative

research.

The shininess of a giant n may be alluring to researchers who want to run regressions with dozens of independent (Roman 2023) or dependent (Fraga et al. 2025) variables, careful thought needs to be put into the descriptive characteristics of the data used. A dataset should, in theory, be a random sample of the population to make inferences about the population possible. Representativeness should be understood as a sample's respondents approximately matching the population across various demographics (Grafström and Schelin 2014). Because this is the goal, it is worth imagining what properties a perfect Race and Ethnic Politics dataset would have.

It's hard to say what an ideal dataset for Race and Ethnic Politics would look like. Those behind the CMPS have tried their best, but their randomness and sampling are not as strong as more established surveys (demonstrated in descriptive section) such as the ANES or CES. The CES' n-size is an enticing pull for Race and Ethnic Politics scholars, and an oversample could be less relevant when 60,000 respondents are involved. The ANES asks more questions than the CES and has feeling thermometers. I think that in an ideal world, the infrastructure of the CES could be leveraged with better sampling procedures, more measures, and occasional oversamples of minorities. All three of these major election studies have benefits and drawbacks, so it is important to weigh which of these pros and cons matter the most for Race and Ethnic Politics work.

Thus the question for Race and Ethnic Politics researchers is one of benefits and costs. Which dataset maximizes validity and reliability for any given subsample?

I think that the deficiencies in the CES will make it perform poorly compared to the CMPS for racial subsamples. First, the lack of questionnaires in non-English languages seriously complicates the descriptive universe for minority subsamples, especially Hispanics and Asians. As sufficient evidence in the literature has shown, language has a serious effect on politics that cannot be taken lightly. Ignoring the non-English-speaking segment of the population is both ethically questionable and scientifically dubious. Lopping off a possible class of respondents based on a demographic characteristic makes the CES less representative than it can be, not only for racial minorities, but for any analysis on American Politics, generally, especially in states with large non-English-speaking populations. The CES also claims to have representative samples in all Congressional Districts, but some of those districts have Hispanic populations of up to 90% (Texas Legislature 2023), so any small errors in validity would be inflated when doing analyses on specific areas that have large amounts of Spanish Speakers. These two aspects make using the CES highly questionable in Race and Ethnic Politics research.

The ANES, on the other hand, has a smaller sample size but more measures and more complicated weights. The ANES contained two associated subsamples that comprised part of their overall respondent base, the GSS subsample and a panel study started in 2016 (DeBell et al. 2022, 7). All recruiting and interview materials were translated into Spanish (DeBell et al. 2022, 31). A complexing weighing system was used to take into account different subsamples, and weights were introduced for interactions of age, gender, race, education, marital status, census region, and standalone terms were included for nativity, population density, income, early voter

status, and vote choice in 2016 (DeBell et al. 2022, 86). I think that in terms of the way the sample was taken, the availability of Spanish language interview materials, and the weighting scheme, the ANES is the best possible dataset for use in Race and Ethnic Politics from a methodological standpoint. The only problem lies in its sample size for minority groups.

I think that both descriptive statistics and regression outputs will be deeply affected by sampling patterns of minority groups for subsamples of each dataset. Given the CES' lack of Spanish respondents, I expect the CES specifically to be different from all other studies in this paper, which do offer surveys in Spanish. I also think that the CMPS coefficients will be closest to those in the ACS for minority subsamples only.

4 Research Design

The purpose of this paper is to better understand the implications and biases introduced by the idiosyncrasies of certain American politics datasets when using them to analyze subnational opinion.

A technical report by Deshpande and Cha (2022) examines discrepancies in the distribution of demographic characteristics for Hispanic and Asian respondents in the 2020 CES as compared to the 2019 ACS. They find that the CES undersamples people of lower education, Hispanics in the West, Indians, and Mexicans (Deshpande and Cha 2022, 2). Oddly enough, 17.6% of the CES' Hispanic sample claims to be from or descended from Spain (Deshpande and Cha 2022, 3), though this may be a result of people claiming to be solely Spanish. Asian subpopulations have too small

a sample size for disaggregated analyses by ethnic group, the same is true for voter-validated Hispanics (Deshpande and Cha 2022, 4). Generation and immigration status for Asians and non-Mexican Hispanics are also not sufficiently sized to allow any meaningful analysis (Deshpande and Cha 2022, 5). The CES also problematically allows Asians to select a "United States" origin in place of giving an answer that would identify ancestry or ethnic group (Deshpande and Cha 2022, 10).

4.1 Descriptive Framework

Given that there has not been a comprehensive design created in the way of evaluating subsamples of datasets in the American Race and Ethnic Politics context, I visit this issue *de novo*. I expand on Deshpande and Cha's (2022) analysis by including several dimensions that matter more for Race and Ethnic Politics work than for American politics work.

1. Language alone Given that racial minorities are more likely to speak minority languages, and their politics are often affected by the use of non-English languages, it is worth evaluating a possible bias introduced by either not offering questionnaires in multiple languages or not offering questionnaires in a sufficient variety of languages for the subgroup meant to be analyzed. This is evaluated against benchmarks in the PUMS data.

2. Language by geography The common usage of non-English languages is not evenly distributed (as will be shown in exploratory data analysis of the PUMS) and can bias estimates from particular areas.

3. Ethnicity Continuing from the analysis of Deshpande and Cha (2022), and

considering that Hispanics and Asians are heterogeneous groups that have different political behaviors for granular ethnic groups (Wong et al. 2011; Derek Wakefield 2025), ethnicity is measured as a bias in two ways. First, bias from ethnicity is considered in the whole sample of the survey. Second, bias from ethnicity is considered in the context of racial group. What this means is that, for example, the sample of Puerto Ricans will be tested for bias as a percentage of the whole population, and as a percentage of Hispanics. This means that estimates of bias can be used for both general American politics research and research on Hispanics. This is benchmarked against the PUMS.

4. Race by geography Racial minorities are not evenly distributed across America and form distinct pockets in many geographic areas. Given the previously established importance of geography to Race and Ethnic Politics, bias is measured for each racial group's distribution by state or region. This benchmarked against the PUMS.

5. Race by education and 6. Race by income Because education and income are commonly introduced survey weights and because there are disparities in education and income linked to race, it is important to Race and Ethnic Politics research to understand whether certain racial groups are sampled or weighed by their own distributions of education and income, rather than national distributions of education and income.

8. Race by partisanship and 9. Race by ideology These final tests for bias show the differences between American Politics and Race and Ethnic Politics datasets' measurements of political opinions by race in order to demonstrate the

substantive effect of higher level sampling disparities on possible outcome variables. These are weighed against the CMPS.

These biases are calculated against possible benchmarks. Given that this paper analyzes, in part, the relative merits of using American politics datasets rather than smaller n and more infrequent Race and Ethnic Politics datasets, several biases are benchmarked against Race and Ethnic Politics datasets. However, these data sources are not the best benchmarks for demographics at large, and cannot (should not) be used to calculate post-stratification weights. Thus, the main benchmark used is Public Use Microdata Sample (PUMS) estimates from the American Communities Survey (ACS) from 2019 and 2023 (U.S. Census Bureau 2023). These estimates include high-dimensional contingency tables for many variables, including race, ethnicity, state, citizenship, language, and English proficiency. This benchmark is not perfect, and limitations are further expanded upon in the PUMS section. In short, measurements are not as precise as possible. Descriptive statistics for the PUMS for language is also analyzed.

For large n and time series American politics datasets, I analyze the CES and the ANES. I compare these datasets to the CMPS. All of these studies have been used in some Race and Ethnic Politics research published in the last 20 years. I measure for descriptive bias across several domains.

4.2 Correlative Framework

After the descriptive section, I go further than Deshpande and Cha (2022) by testing the outcome of regressions on each dataset. The regressions will include common

explanatory variables, such as race, gender, education, income, and religion, and common dependent variables, like party identification and ideology. Given that the CES and ANES are nationally representative, an identical regression should yield similar results in each year. The same regression is duplicated for Hispanic, Black, and Asian subsamples of the CES and Hispanic and Black subsamples of the ANES to determine how regression analysis of American Politics subsamples differs from oversamples such as the CMPS. Essentially, American Politics datasets should not be vastly different from Race and Ethnic Politics datasets and patterns in people groups in those datasets should remain regardless of what data is used. The alternative would be concerning, as results from Race and Ethnic Politics papers that use American Politics datasets could be contingent on the data used and may not adequately reflect actual patterns in minority communities.

There are two regressions considered that could be reasonably plausible for an average regression one might see in political science research. Regression 1 is

$$income \sim OLS(race + gender + education(ordinal) + age$$

where income is a numeric variable. Because the PUMS has data for income, it is used as the benchmark against which the CES, ANES, and CMPS is measured. Regression 2 is

$$democrat(binary) \sim logit(race+gender+education(ordinal)+age+income+evangelical(binary)$$

and Hispanic subsamples for both replace race with Hispanic ethnicity. All regres-

sions include original weights from the datasets (sample or over-sample weights from CMPS when appropriate). This method of comparing OLS coefficients was suggested to me by Christopher Wlezien in a time-series context and agreed to by David Siegel.

5 Findings

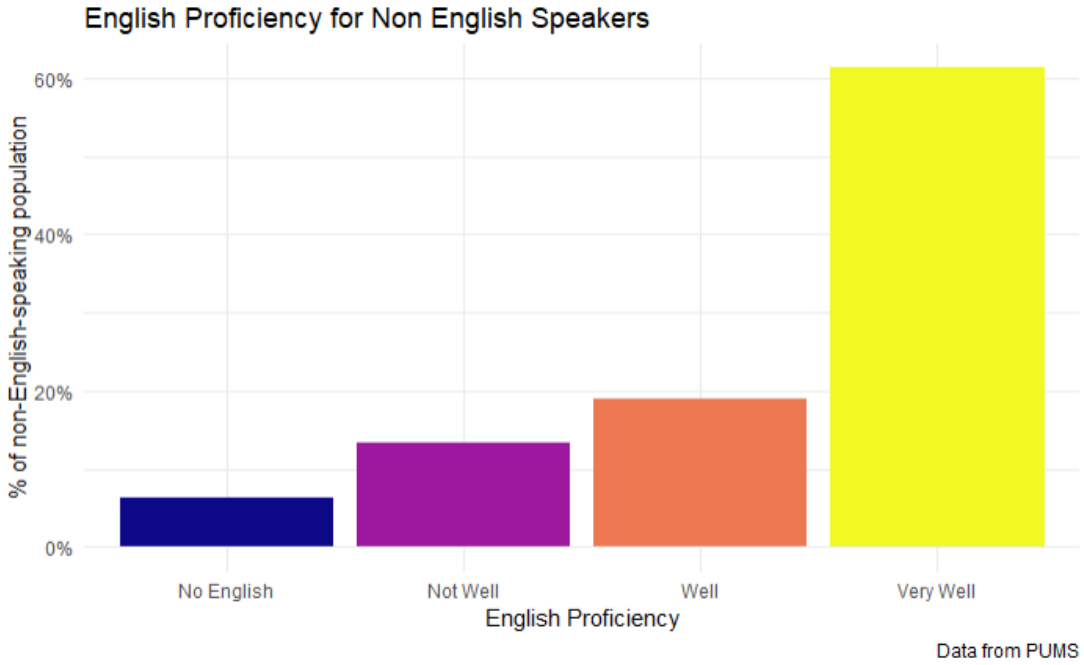
The findings section continues as follows: a brief look at language in the United States as of 2023 through the PUMS, and a discussion of the language bias in datasets. Then I analyze the CES and ANES for points 3 through 9, and the regressions section comparing coefficients from identical regressions in the year 2020.

5.1 PUMS

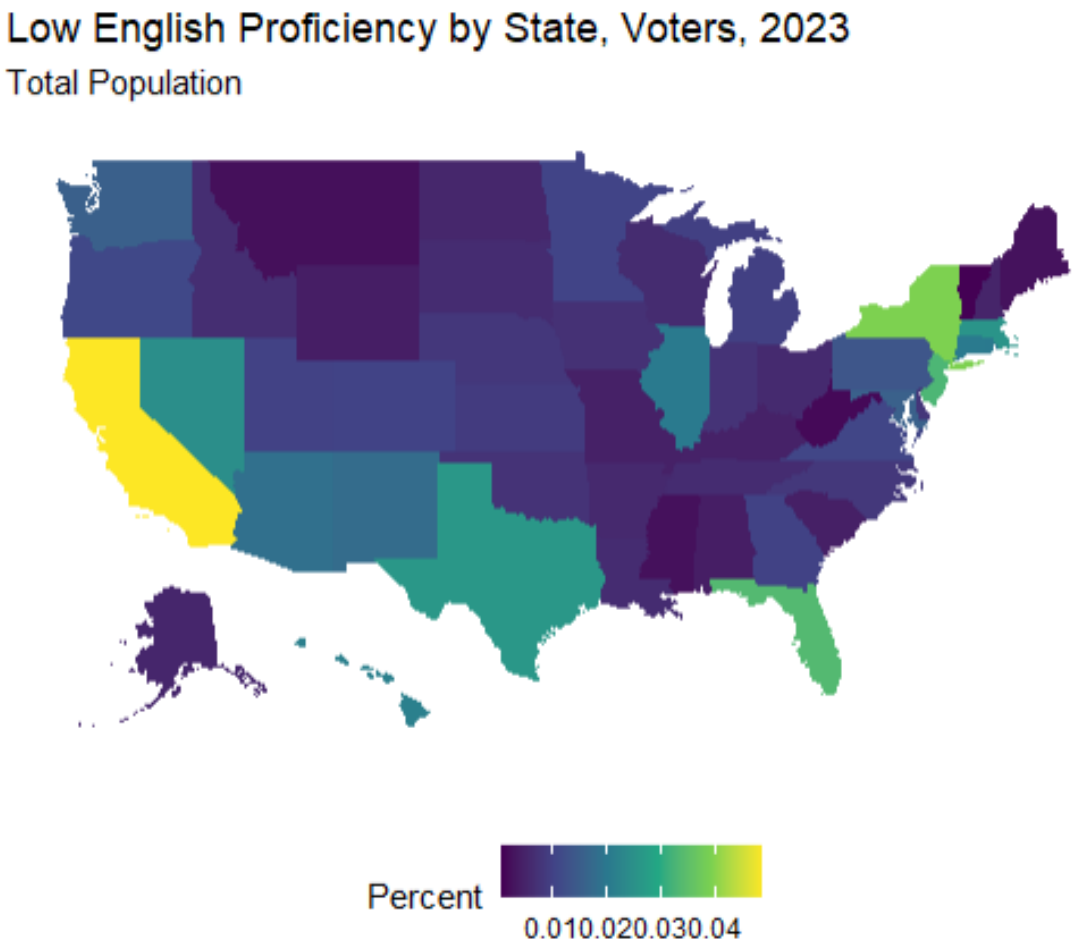
Starting from a high-dimensional crosstab of all relevant variables (cells = 45,883,266) representing 5,696,094,969 estimated observations across 18 years, the PUMS microdata released from yearly ACS surveys is the most comprehensive benchmark possible, and is released by the U.S. government. Because it forms a complete contingency table, it can be used for the creation of post-stratification weights with precise interactions. This allows for analyses of multidimensional biases across datasets.

5.1.1 Language

There are several different ways of examining language in America. Given that American politics revolve around English, the first consideration is the English proficiency of foreign language speakers. The proportion of foreign language speakers who speak English "very well" has increased in recent years, and the proportion that



do not speak English well has decreased. Around 80% of foreign language speakers in America speak English well or very well. At present, less than 5% of voters do not speak English well or very well. The number of people whose household language is Spanish had increased modestly from 15% to 17% in a 20-year span. The proportion whose Household Language is one of many Asian languages is at present 6%. Of these, Indian Subcontinent Languages are the most common, followed by Chinese Languages, Filipino Languages, Vietnamese, Korean, and Japanese. In the wider time series, there was a change in how the household language question was asked to allow for more granularity. This disparity is shown in the figure.


Low English Proficiency by State, Voters, 2023
Total Population
Percent
0.010.020.030.04

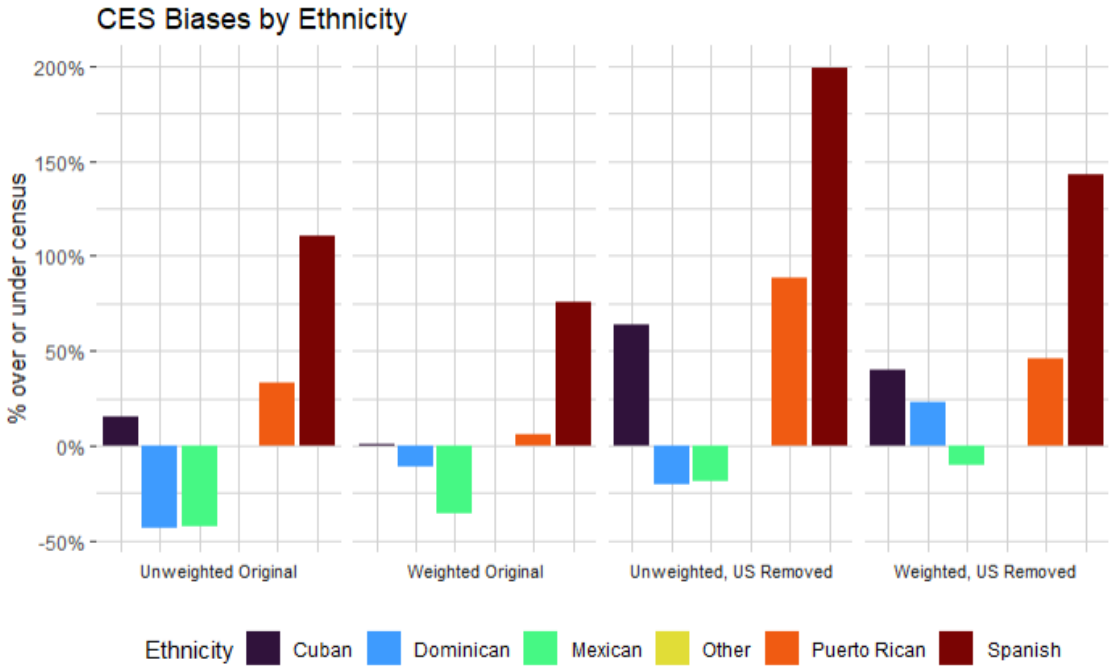
5.1.2 Language by Geography

California, Texas, Florida, New York, and New Jersey have the highest non-English speaking populations. California also has the highest percentage of non-English speaking voters, at 5%. This number is smaller than I initially thought, and it seems that the segment of the population left out of the sample is smaller than I imagined. Additionally, 95% of the respondents in the CMPS took the survey in English, and 99% of the respondents in the ANES took the survey in English. This casts serious doubt on whether language is actually a concern when using the CES. However, many Hispanic respondents claim to speak Spanish often at home. The takeaway is that Spanish language remains important for Hispanic Households, but that there is no serious bias introduced into the CES as compared to the CMPS from not offering non-English surveys. However, the CES' sample of Hispanics is still problematic.

5.2 CES

The main takeaway of this section is that the CES chronically undersamples Hispanics and oversamples Whites. This disparity is so significant that it affects not only Race and Ethnic Politics work, but also any analyses of politics in states with high Hispanic populations or on congressional districts with high Hispanic populations. The CES does remarkably badly at finding Hispanic respondents. They also do not field any surveys in Spanish, so no further analyses can be done with regards to language. Because no language information is available in the CES, I will start the analysis at item 3, ethnicity.

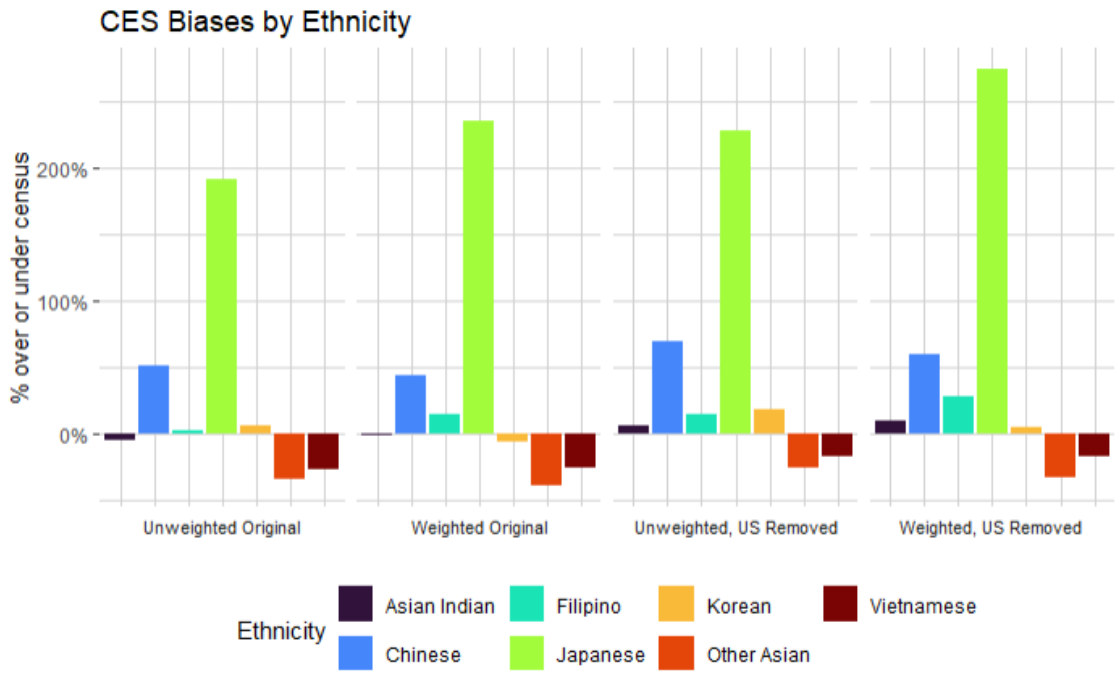
Throughout the following sections, percent bias is defined as $\frac{\%Survey}{\%Benchmark} - 1$ so



that less than the benchmark is negative and greater than the benchmark is positive.

Ethnicity

Both the Hispanic origin and Asian origin questions on the CES have problems. The primary problem, noted by (Deshpande and Cha 2022), is the offering of "United States" as an option for ancestry in both the Hispanic and Asian ethnicity question. This leaves the dataset with Hispanic and Asian respondents who only claim to have ancestry from the United States and with no marker of ethnic group. The census does not allow this option. A weighted 9.3% of Asians and 27.8% of Hispanics choose this option instead of reporting their ethnic group. To correct for this, I consider weighted percentages against the census both with and without these United States



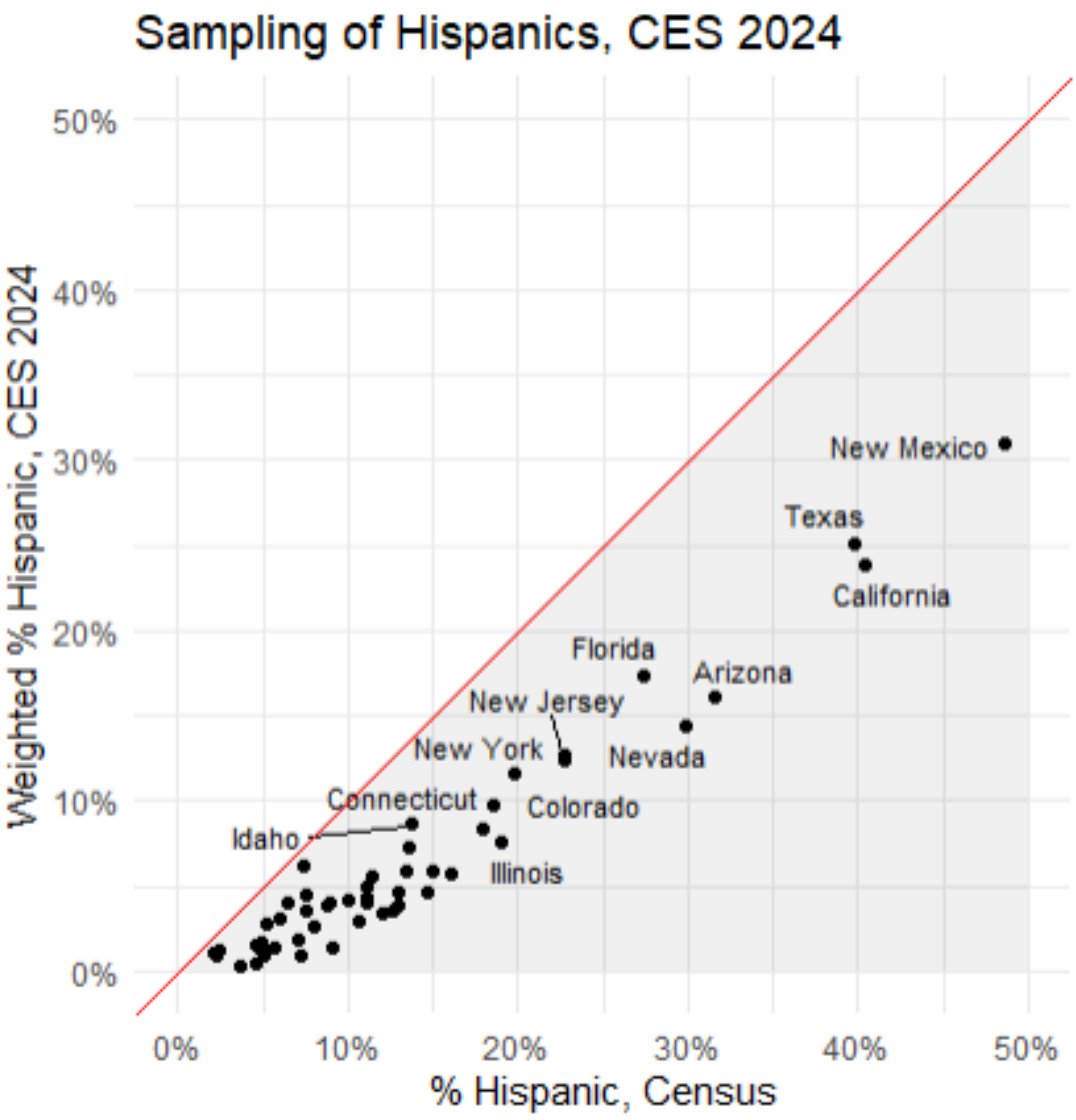
identifiers. Another significant difference between ethnicity questions in the CES and Census is tracking of the "Spanish" population. The Census requires someone to identify as "Spaniard" on the Hispanic ancestry question, whereas the CES asks where one can trace one's ancestry from. Thus, the Census records 2.5% of the Hispanic population as Spanish, whereas the CES records a weighted 4.5%. In the weighted original dataset, Mexicans are undersampled at 43% less than the Census as a percentage of the Hispanic population. When removing United States identifying respondents, Mexicans remain mildly underrepresented and Cubans and Puerto Ricans are overrepresented.

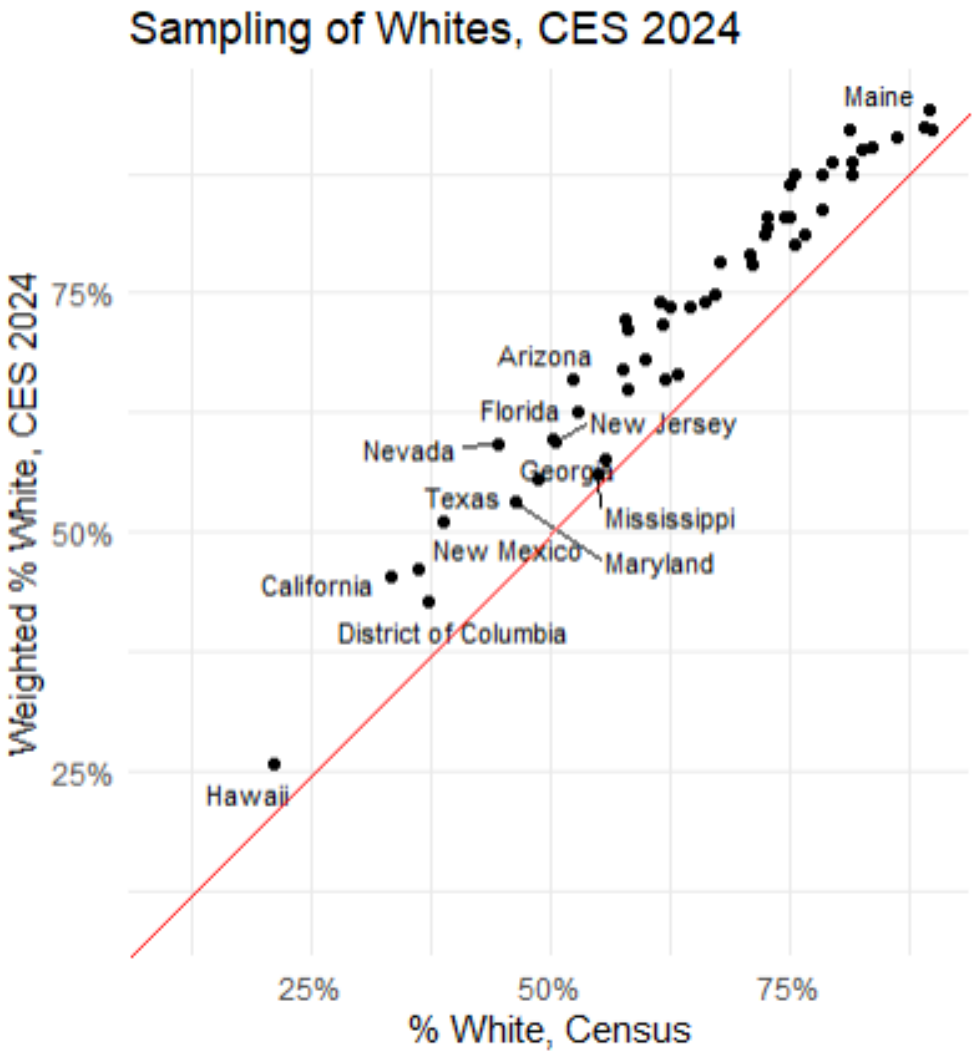
For Asians, Japanese are massively overrepresented in the data, with 234% more than the benchmark from the census. Chinese and Filipino respondents are also over-

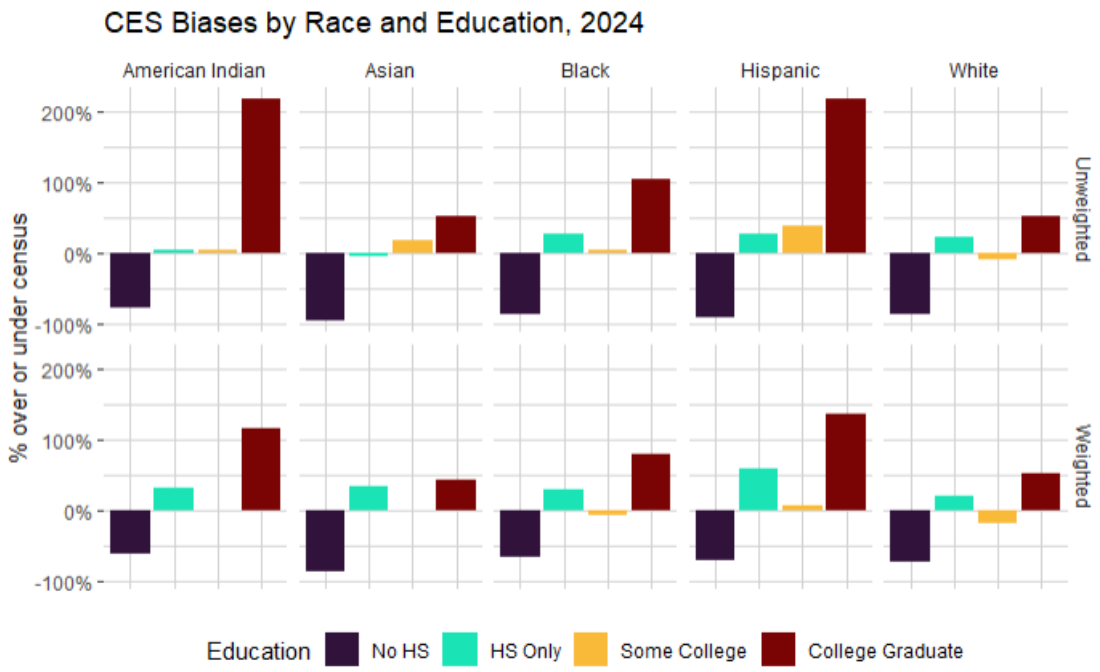
represented. Vietnamese respondents are underrepresented with 25% less than the expected population from the Census in the original dataset and 17% when removing United States respondents. Other Asians are also underrepresented at 39% and 32% less than the Census, respectively. Given these biases, using CES data for any analyses that disaggregate Hispanics or Asians by ethnic group is questionable. Several groups are heavily overrepresented and many are underrepresented. The addition of the United States option makes understanding subgroup political behavior more difficult and likely invalid. Put simple, the CES does not sample Hispanic or Asian subgroups well, and the CES should not be used as part of an analysis of Hispanic or Asian subgroups.

Race by State

This is the most important finding from the CES analysis. The CES deeply undersamples Hispanics, even when weighted. Hispanics are undersampled in every state, and the worst undersamples come from states with high Hispanic populations. Given that the CES claims to have a representative sample in every state, underestimating the Hispanic population of every state, especially New Mexico, 18% off; and Texas, California, Arizona, and Nevada, 15% off, seriously questions whether a sufficient sample exists in every state for analyses of state politics. As will be demonstrated at the end of this section, this has ramifications for American Politics research as well as Race and Ethnic Politics. Blacks are sampled reasonably consistently with the Census. Asians are sampled below the benchmark in most states. Though this does not deeply bias large-scale analyses, the CES sample of Asians is inconsistent across


Sampling of Hispanics, CES 2024
Weighted % Hispanic, CES 2024
% Hispanic, Census
0%
10%
20%
30%
40%
50%
New Mexico
Texas
California
Florida
Arizona
New Jersey
New York
Nevada
Connecticut
Colorado
Idaho
Illinois


Sampling of Whites, CES 2024
Maine
Arizona
Florida
New Jersey
Nevada
Georgia
Texas
Mississippi
New Mexico
Maryland
California
District of Columbia
Hawaii
75%
50%
25%
Weighted % White, CES 2024
25%
50%
75%
% White, Census



state, as well as ethnicity. Another important aspect to consider is the oversampling of Whites. Just as Hispanics are undersampled in every state, Whites are oversampled in every state. This oversample is smaller in magnitude than the undersample of Hispanics, but compounds with that undersample in Texas, California, and New Mexico. The RMSE for Whites is 8.7% and for Hispanics is 8%. Despite having a large n of racial groups, the CES deeply fails at sampling Hispanics in the right places and is biased in favor of Whites. Descriptive work using the CES, even in an American politics context will be flawed because of this disparity.

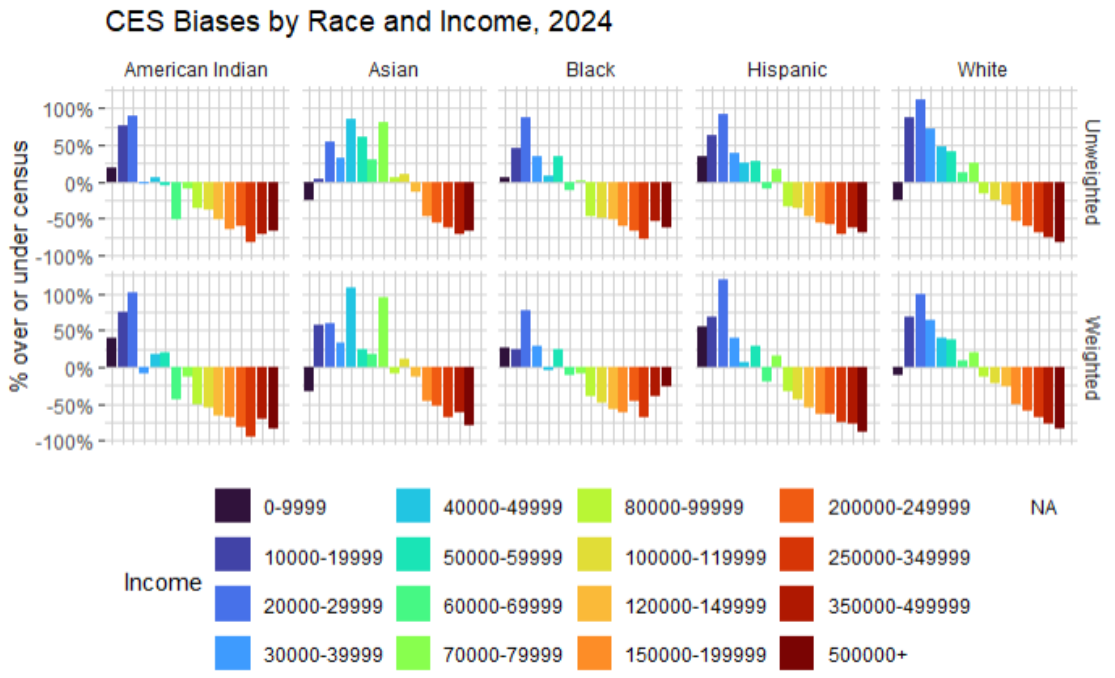
5.2.1 Race by Education

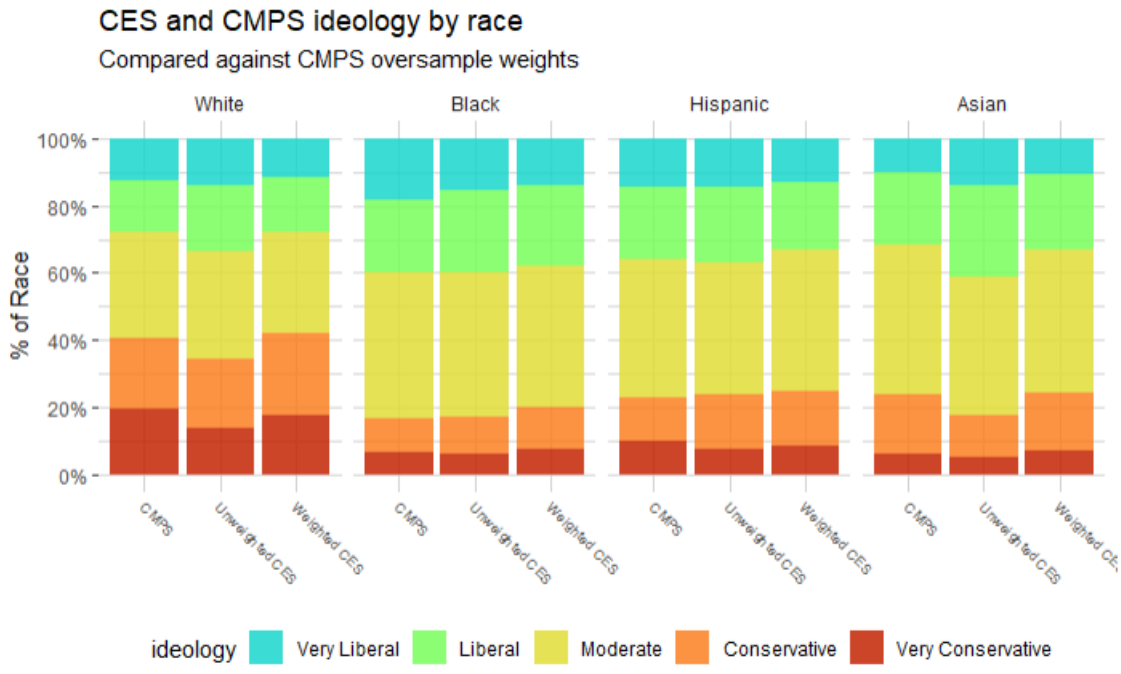
Unsurprisingly, higher educated people are overrepresented in the sample across all racial groups. However, this overrepresentation is most drastic with the Hispanic

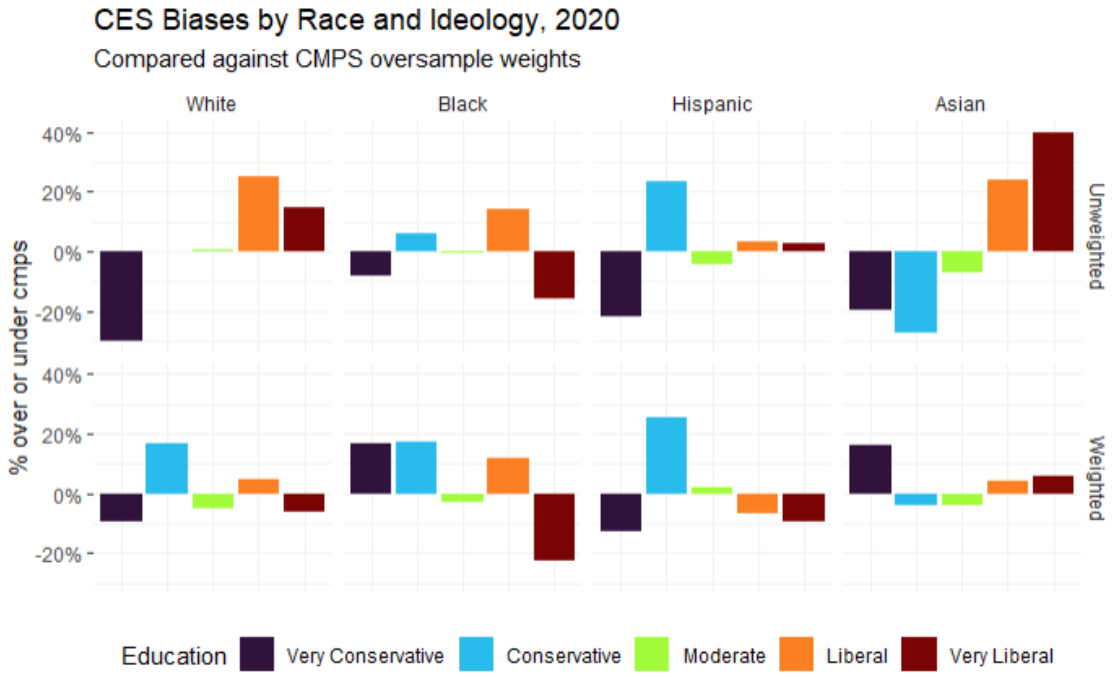
subsample. There are about three times as many college graduates as there should be. High school graduates are also overrepresented. Non-High school graduates are underrepresented across all races, but the disparity is largest for Asians. The subsamples for Black, Hispanic, and Asian respondents are more biased across the board than the subsample for White respondents. Given the wide difference between the weighted and unweighted biases, the survey weights likely incorporate education, and those weights are doing heavy lifting. However, those weights do not adequately fix the deep sampling bias and post-weighting biases are especially prominent among minorities, especially Hispanics. Minority subsamples of the CES will be deeply biased by education, even when taking weights into account.

Race by Income

As might be expected by the design of surveys, richer people are underrepresented in the data and poorer people are overrepresented. Given that minorities are poorer than Whites, these sampling biases lead to the unusual effect of the data being more representative of some minorities than of Whites. The Black subsample has the least overall bias, followed by Asians, Whites, and Hispanics. However, negative sampling biases for Black and Hispanics appear earlier, with those in the $80,000 to $149,999 range being far more underrepresented than for the White or Asian samples. Thus, there are fewer observations of Upper Middle Class Blacks and Hispanics, which may complicate intersectional analyses of class and race.


CES Biases by Race and Income, 2024
American Indian
Asian
Black
Hispanic
White
Unweighted
Weighted
% over or under census
100%
50%
0%
-50%
-100%
Income
0-9999
10000-19999
20000-29999
30000-39999
40000-49999
50000-59999
60000-69999
70000-79999
80000-99999
100000-119999
120000-149999
150000-199999
200000-249999
250000-349999
350000-499999
500000+
NA


CES and CMPS ideology by race
Compared against CMPS oversample weights
White
Black
Hispanic
Asian
% of Race
100%
80%
60%
40%
20%
0%
CMPS
Unweighted CES
Weighted CES
ideology
Very Liberal
Liberal
Moderate
Conservative
Very Conservative

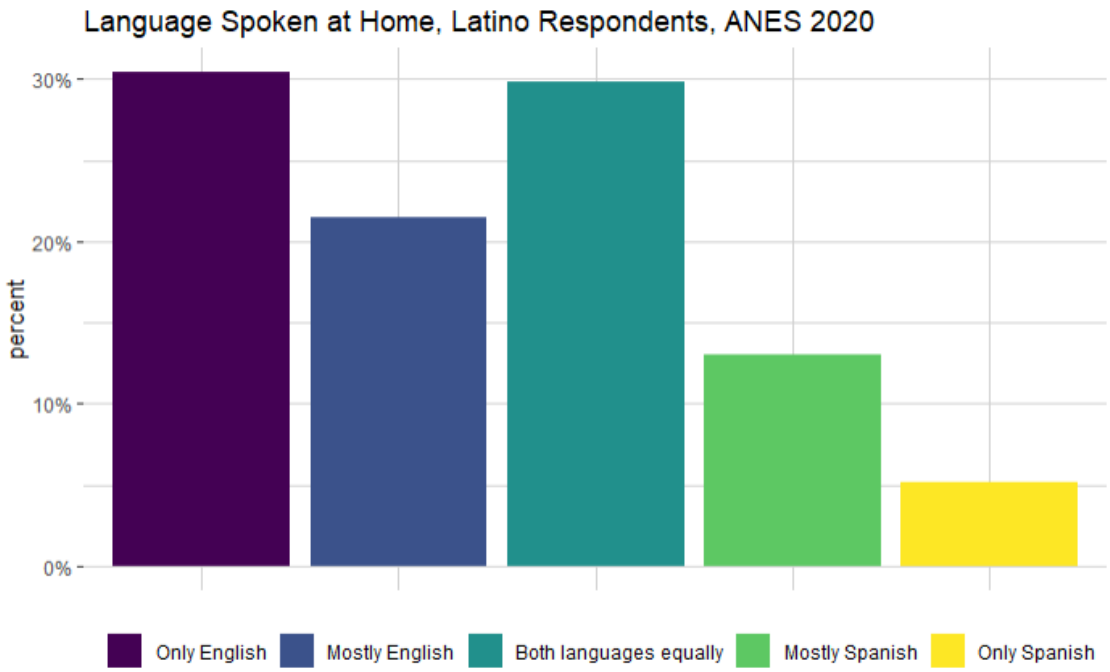


Race by Ideology

In the weighted CES, conservative Black and Hispanic respondents are marginally overrepresented and liberal Hispanics are marginally underrepresented. However, these differences are very small.

Race by Party ID

For race by party ID and ideology, the CES 2020 is compared against the CMPS 2020. The weighted CES is very similar to the CMPS in terms of Black, Hispanic, and Asian Republicans. However, there are slightly fewer Black and Hispanic Democrats in the CES than the CMPS and more Independents and "Others", likely an artifact of how the questions were worded. The CES had "Other Party" as an option and



the CMPS did not.

Reflections on the CES

The CES has far too many Whites and too few Hispanics, especially in highly Hispanic states. This difference is likely not caused by language. Further biases are seen in ethnicity breakdowns for Hispanics and Asians, education for Hispanics, Blacks and Asians, and income for Upper Middle Class Blacks and Hispanics. However, no serious deviation from the CMPS for ideology or party ID is found.

5.3 ANES

Language

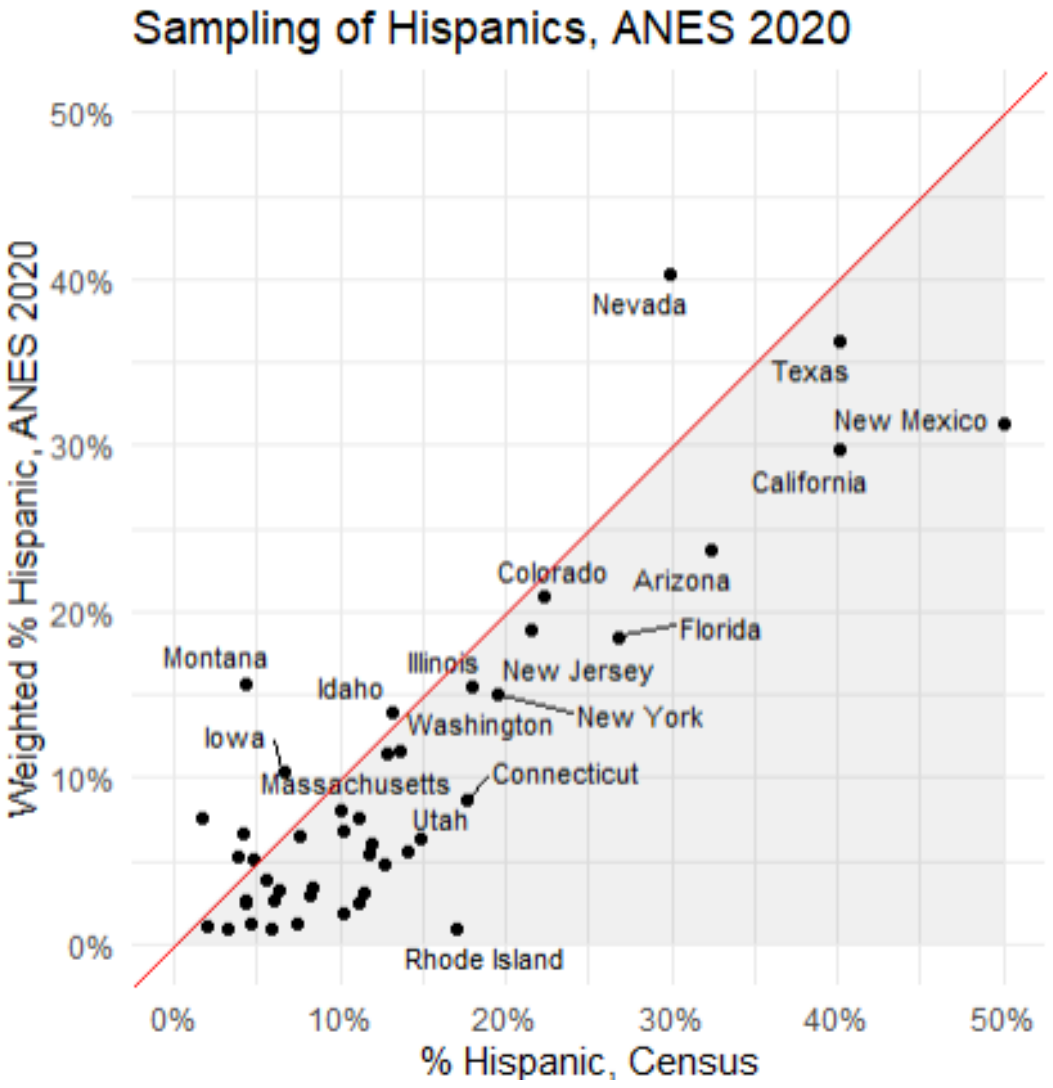
The ANES does have household language and interview language available, but these are on different scales than the PUMS, so they cannot be compared easily. Specifically, the PUMS only asks for the household language, whereas the ANES asks for how much of a household language is spoken. The ANES does not ask questions about English proficiency. Less than 1% of ANES respondents were interviewed in Spanish. However, the majority of Hispanic respondents said that they spoke Spanish at least the same amount as English in their household.

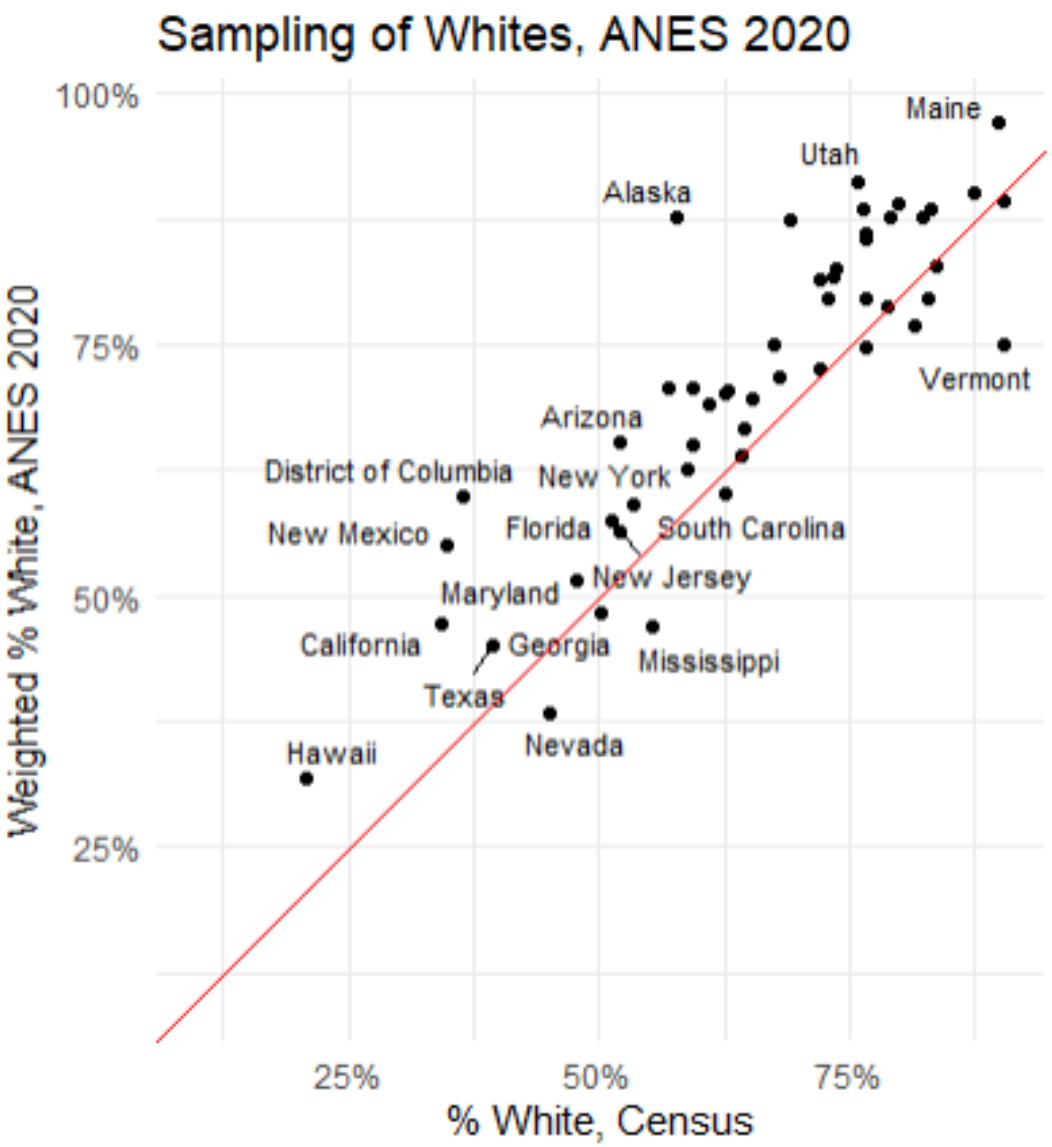
Ethnicity

The ANES has far less granularity with their Hispanic ethnicity question and does not have an Asian ethnicity question. The only options for a respondent are "Mexican", "Puerto Rican", or "Other Hispanic". The unweighted ANES has 10% too few Mexicans and the weighted ANES has 5% too few. This is far better than the CES.

Race by State

The ANES also undersamples Hispanics in most states, but not to the same extent as the CES. The RMSE for Hispanics is 6%. Similarly to the CES, the ANES oversamples Whites, with an RMSE of 9%, and samples for Blacks and Asians across states are more varied but are approximately symmetric for oversamples and undersamples. New Mexico remains the most incorrectly predicted state, with the ANES


Sampling of Hispanics, ANES 2020
Weighted % Hispanic, ANES 2020
% Hispanic, Census
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
Nevada
Texas
New Mexico
California
Colorado
Arizona
Florida
Montana
Illinois
New Jersey
Idaho
New York
Washington
Iowa
Massachusetts
Connecticut
Utah
Rhode Island


Sampling of Whites, ANES 2020
Weighted % White, ANES 2020
% White, Census
100%
75%
50%
25%
25%
50%
75%
Maine
Utah
Alaska
Vermont
Arizona
District of Columbia
New York
New Mexico
Florida
South Carolina
New Jersey
Maryland
California
Georgia
Mississippi
Texas
Nevada
Hawaii

being 20% off of the Census Hispanic population.

Race by Education

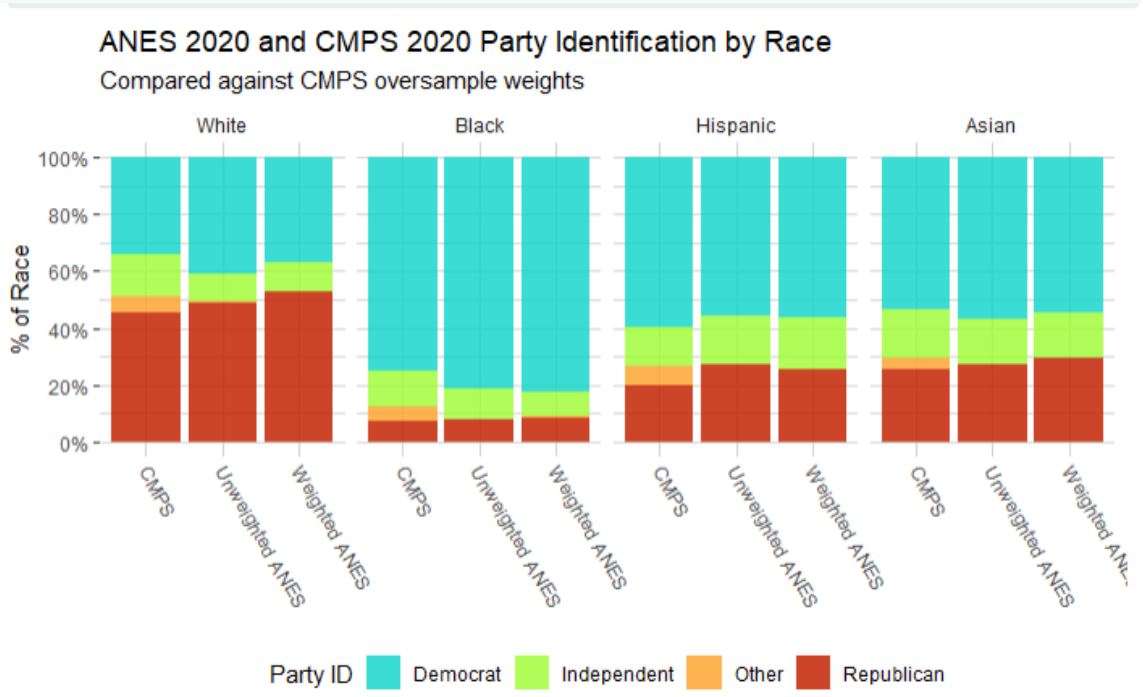
As may be expected, people with higher education are overrepresented in the ANES sample. This disparity is greatest for Hispanics, followed by Asians and Blacks. All three racial groups have greater biases than the White sample. It is also remarkable how far off the original sample was, as college graduates were 150% more than the Census and "some college" was 100% more than the Census. The same bias for education by race is seen in the ANES.

Race by Income

The ANES bias by race and income is more evenly distributed than that of the CES, and the survey weights do a lot in changing the distributions of bias. It seems that Asians have the largest biases: there are far too few low income Asians and too many middle income Asians. Hispanic and Black biases are less significant.

Race by Ideology

The CMPS has more people on the extremes of the ideological Likert scale, whereas the ANES has far less strong Conservatives and Liberals. The ANES has almost no "Very Conservative" Black, Hispanic, Asian, or White respondents than the CMPS. This should be taken into consideration if ideology is involved as a variable in a regression.

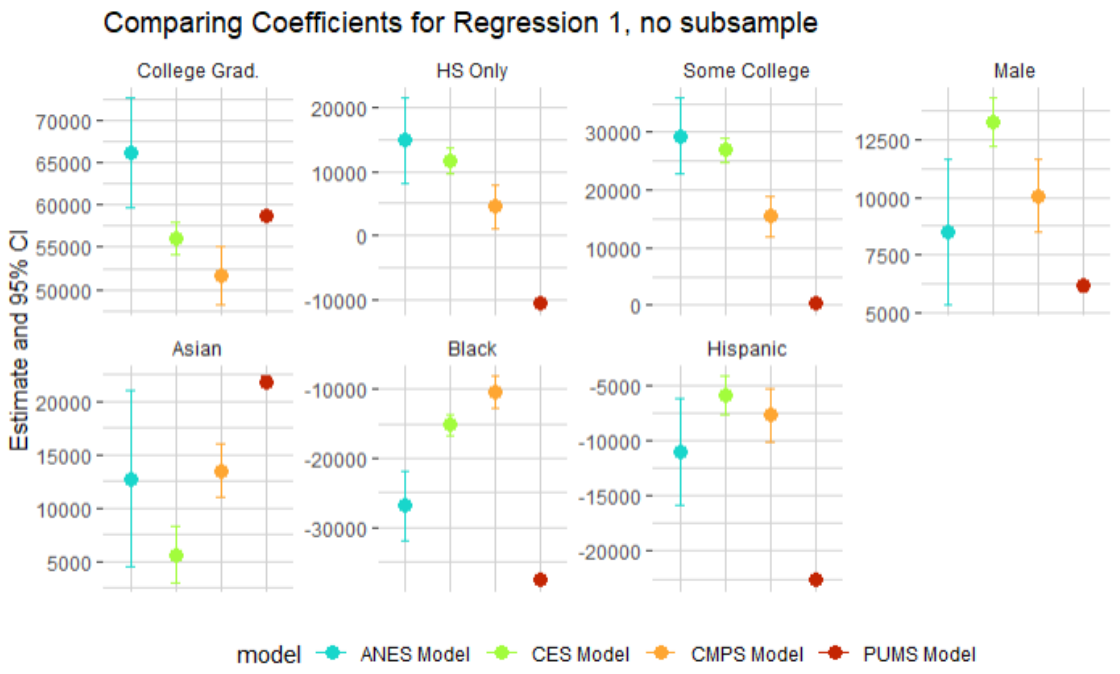


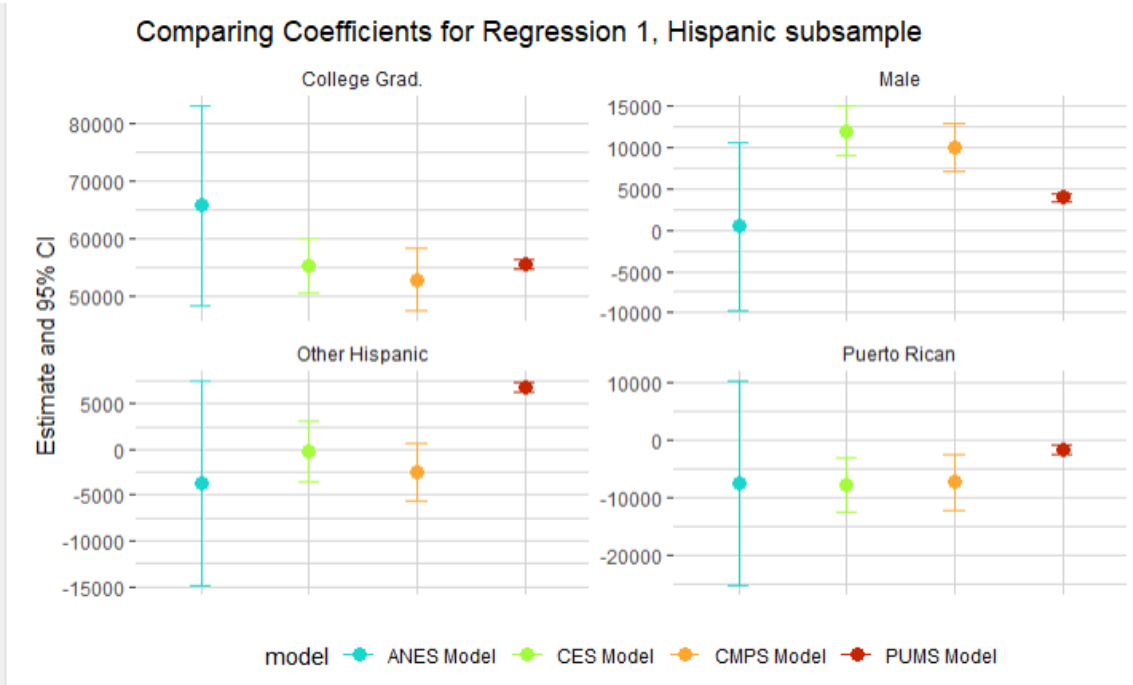
Race by Party ID

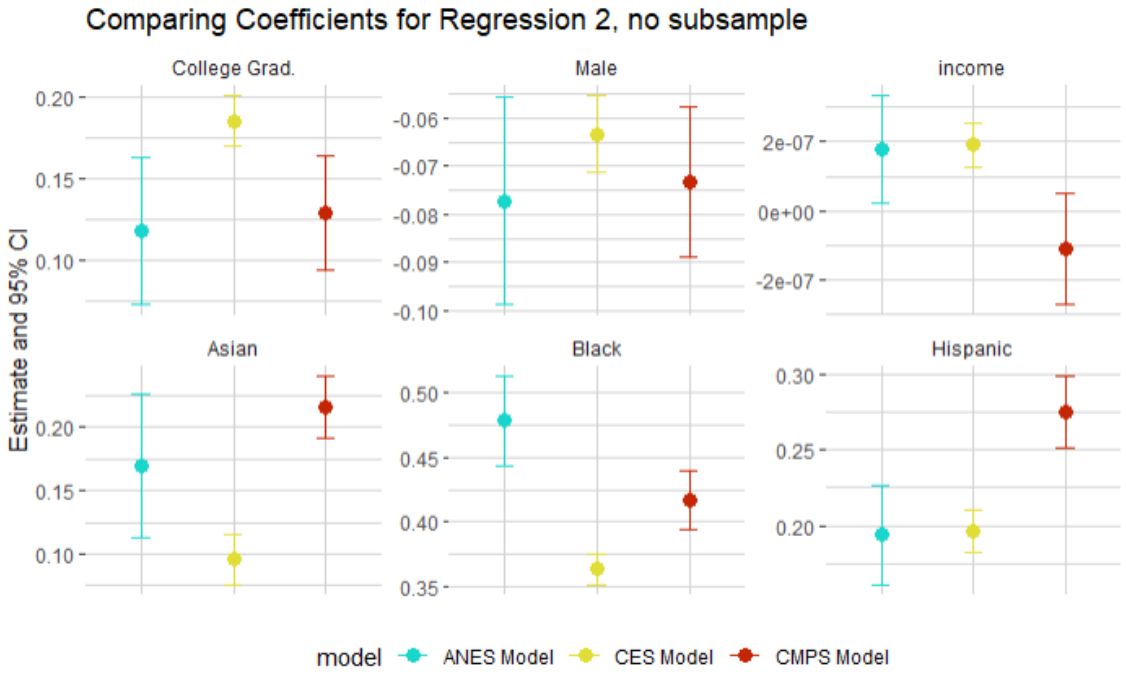
The ANES codes party ID more conservatively than the CES or CMPS, and does not allow an easy answer for an "Other" Category. This seems to take from Hispanic and Asian Republicans, whereas the Black ANES sample is more Democrat than the Black CMPS sample.

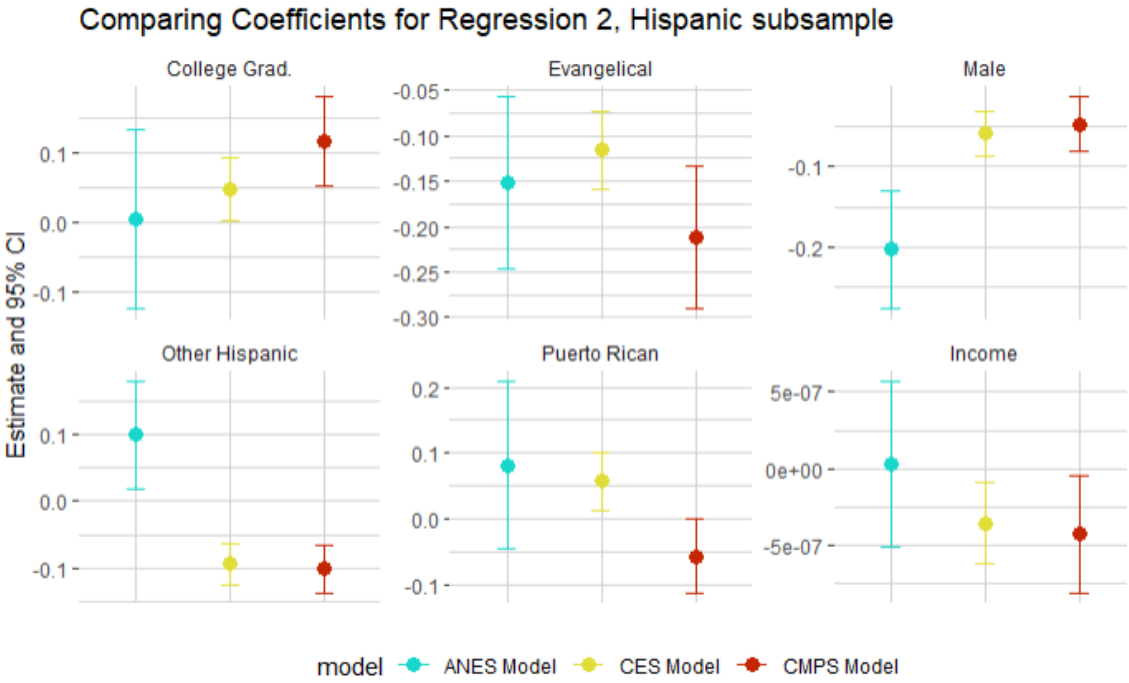
Reflections on the ANES

The ANES has much of the same biases as the CES, though usually to a lesser extent. Though they undersample Hispanics, they do so to a lesser extent.


Comparing Coefficients for Regression 1, no subsample
College Grad.
HS Only
Some College
Male
Asian
Black
Hispanic
Estimate and 95% CI
model
ANES Model
CES Model
CMPS Model
PUMS Model


Comparing Coefficients for Regression 1, Hispanic subsample
College Grad.
Male
Other Hispanic
Puerto Rican
Estimate and 95% CI
model
ANES Model
CES Model
CMPS Model
PUMS Model


Comparing Coefficients for Regression 2, no subsample
College Grad.
Male
income
Asian
Black
Hispanic
Estimate and 95% CI
model
ANES Model
CES Model
CMPS Model


Comparing Coefficients for Regression 2, Hispanic subsample
College Grad.
Evangelical
Male
Other Hispanic
Puerto Rican
Income
Estimate and 95% CI
model
ANES Model
CES Model
CMPS Model

Regressions

This section includes identical regressions run on the PUMS, CES, ANES, and CMPS, from 2020 and 2021. These regressions should return similar coefficients regardless of what dataset is used, given the latent patterns in the population at large.

As a reminder, regression 1 is

$$income \sim OLS(race + gender + education(ordinal) + age$$

and Regression 2 is

$$democrat(binary) \sim logit(race+gender+education(ordinal)+age+income+evangelical(binary)$$

For the numeric outcome of income, all three datasets are closer to each other than they are to the PUMS. Especially notable are the education variables; the intercept for the PUMS is far higher than the datasets, with the same reference categories across datasets. I don't know why this is the case, perhaps something to do with the Census' weighing. However, the ANES and CES predict education having a far larger coefficient on income than the CMPS.

For the Hispanic subsample of regression 1, there are no significant differences between datasets.

For regression two, democratic identification, the CES and ANES coefficients on Hispanic are significantly less than the CMPS' which means that the CES and

ANES underestimate the coefficient of being Hispanic on being a Democrat. The CES also significantly underestimates the same effect for Black and Asian respondents. In a regression where race is an explanatory variable, The CES underestimates the importance of race compared to the CES. The CES also underestimates the coefficient of being a college graduate on being a Democrat. In the Hispanic subsample, the only significant difference between datasets is between the CES and CMPS which predict opposite coefficients of being Puerto Rican on Democratic identification. For the Black and Asian subsamples, there is not significant differences between datasets.

This suggests a very interesting finding. When estimating the coefficient of a race on a whole population, American Politics datasets are significantly different from Race and Ethnic Politics datasets in that they underestimate effects for race. But when subsamples of American Politics datasets are used, those subsamples are not significantly different than Race and Ethnic Politics datasets in regression models.

6 Discussion

I think that the findings of this paper have interesting implications for how we think about doing observational work in Race and Ethnic Politics. A collected sum of the findings of this paper are that: the CES and ANES undersample Hispanics, this effect is gigantic for the CES. Language of interviews does not appear to have an effect on the samples collected. There are various interactional biases such as race by education present in the data. Regression coefficients for race in a regression run on a whole population will be smaller in magnitude for American Politics datasets than for the CMPS. However, subsamples of American Politics datasets by race are

not significantly different from subsamples of the CMPS by race.

Language is not the culprit of the CES' Hispanic problem. But something else must be. They have far too few Hispanic observations from every state, especially states with large Hispanic populations. I remain somewhat skeptical of the use of American Politics subsamples in Race and Ethnic Politics, but their use is not significantly worse than using the CMPS. The likely limitation of this paper is the lack of a sophisticated benchmark that is more valid than the CMPS. Additionally, the lack of common language variables made understanding the role of language in American Politics datasets rather difficult, though very few CMPS respondents used non-English options, either.

Geography, interacted with race, is the new problem for the CES if they want to claim a representative sample in every state. Their weighted population distributions are vastly different from the Census. Given that the CES is contingent on having a reliable geographically stratified sample, irregularities at the state level call into question the reliability of measurements at the district level, especially given that errors in smaller subsamples compound (Schaffner et al. 2025, 18). However, on balance, it is not too bad of a practice to use American Politics datasets for Race and Ethnic Politics work, keeping in mind the biases in the underlying sample involved.

7 Conclusion

At the beginning of this paper, I sought to understand why Hispanic politics scholars were using a dataset that included no Spanish-speaking respondents (Fraga et al. 2025). I have come to find that language is not a major factor involved. Rather, sys-

tematic disparities in representation of Hispanics in both the CES and ANES affect how regressions work when Hispanic is included as an explanatory variable. However, subsamples of the CES and ANES are not significantly worse than the CMPS, even when using original survey weights. It may not be all too necessary to reweigh (Fraga et al. 2025) or post-stratify (Derek Wakefield 2025) these subsamples when running basic regressions. It seems that the practice of using American Politics data is here to stay. However, Race and Ethnic Politics scholars and practitioners should remain mindful of what descriptive and correlative biases exist in the underlying data.

Figures made using viridis color scales for visual disabilities (Garnier et al. 2024).

References

Abrajano, Marisa and Costas Panagopoulos. 2011. "Does language matter? the impact of spanish versus english-language gotv efforts on latino turnout". *American Politics Research* 39 (4): 643–663.

American National Election Studies. 2022. "ANES Time Series Cumulative Data File [dataset and documentation]. September 16, 2022".

Ansolabehere, Stephen, Samantha Luks, and Brian F Schaffner. 2015. "The perils of cherry picking low frequency events in large sample surveys". *Electoral Studies* 40 : 409–410.

Barreto, Matt A, Fernando Guerra, Mara Marks, Stephen A Nuño, and Nathan D Woods. 2006. "Controversies in exit polling: Implementing a racially stratified homogenous precinct approach". *PS: Political Science & Politics* 39 (3): 477–483.

Barreto, Matt A, Tyler Reny, and Bryan Wilcox-Archuleta. 2017. "Survey methodology and the latina/o vote: Why a bilingual, bicultural, latino-centered approach matters". *Aztlan: A Journal of Chicano Studies* 42 (2): 211–227.

Benjamin, Andrea and Sydney L Carr. 2022. "Does incumbency matter? black voter support for non-incumbent poc democratic candidates in the 2018 congressional house of representative elections". *National Review of Black Politics* 3 (1-2): 2–16.

Brady, Henry E and David Collier. 2010. *Rethinking social inquiry: Diverse tools, shared standards*. Rowman & Littlefield Publishers.

Brown, Trevor E, Gisela Pedroza Jauregui, Suzanne Mettler, and Marissa Rivera. 2025. "A rural-urban political divide among whom? race, ethnicity, and political behavior across place". *Politics, Groups, and Identities* 13 (1): 229–242.

Chan, Stephanie. 2022. *Creative Citizenship: The Impacts of Racialized Incorporation on Political Participation*. PhD diss., Princeton University.

Chong, Chinbo. 2019. *Impact of Pan-Ethnic Identity Appeals on Asian American and Latino Political Behavior*.

Corral, Álvaro J and David L Leal. 2024. "When demography is (not) destiny: Exploring identity and issue cross-pressures among latino voters in the 2020 presidential election". *Social Science Quarterly* 105 (4): 1239–1252.

Cramer, Katherine J. 2016. *The politics of resentment: Rural consciousness in Wisconsin and the rise of Scott Walker*. University of Chicago Press.

Cuevas-Molina, Ivelisse. 2023. "White racial identity, racial attitudes, and latino partisanship". *Journal of Race, Ethnicity, and Politics* 8 (3): 469–491.

DeBell, Matthew, Michelle Amsbary, Ted Brader, Shelley Brock, Cindy Good, Justin Kamens, Natalya Maisel, and Sarah Pinto. 2022, 08. "Plan c2193". *American National Election Studies*.

Derek Wakefield, Bernard Fraga, Colin Fisk. 2025. "Partisan change with generational turnover: Latino party identification from 1989-2022".

Deshpande, Pia and Jeremiah Cha. 2022. "Technical report: Duplicate respondents + asian and hispanic subsamples". *Cooperative Election Study*.

Enders, Adam M and Judd R Thornton. 2022, 05. "Polarization in black and white: An examination of racial differences in polarization and sorting trends". *Public Opinion Quarterly* 86 (2): 293–316.

Fairdosi, Amir Shawn and Jon C Rogowski. 2015. "Candidate race, partisanship, and political participation: when do black candidates increase black turnout?". *Political Research Quarterly* 68 (2): 337–349.

Foxworth, Raymond, Cheryl Ellenwood, and Laura E Evans. 2025. "Surveying native americans: Early lessons from the cmps". *PS: Political Science & Politics* 58 (2): 328–332.

Fraga, Bernard L., Yamil R. Velez, and Emily A. West. 2025. "Reversion to the mean, or their version of the dream? latino voting in an age of populism". *American Political Science Review* 119 (1): 517–525.

Frank, Andrea. 2003. *Chapter 11, Socio-economic applications of geographic information science / edited by David Kidner, Gary Higgs, and Sean White.* Innovations in GIS ; 9. London ;: Taylor Francis.

Frasure, Lorrie, Janelle Wong, Edward Vargas, and Matt Barreto. 2024. "Collaborative multi-racial post-election survey (CMPS), united states, 2020".

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Antônio Pedro, Sciaini, Marco, Scherer, and Cédric. 2024. *viridis(Lite) - Colorblind-Friendly Color Maps for R.* viridis package version 0.6.5.

Geiger, Jessica R and Tyler T Reny. 2024. "Embracing the status hierarchy: How immigration attitudes, prejudice, and sexism shaped non-white support for trump". *Perspectives on Politics* 22 (4): 1015–1030.

Grafström, Anton and Lina Schelin. 2014. "How to select representative samples". *Scandinavian Journal of Statistics* 41 (2): 277–290.

Harris-Lacewell, Melissa Victoria. 2010, 06. *Barbershops, Bibles, and BET*. Princeton University Press.

Hero, Rodney E and Caroline J Tolbert. 1996. "A racial/ethnic diversity interpretation of politics and policy in the states of the us". *American Journal of Political Science*: 851–871.

Hopkins, Daniel J. 2011. "Translating into votes: the electoral impacts of spanish-language ballots". *American Journal of Political Science* 55 (4): 814–830.

Irizarry, Giovanni Castro. 2024. "The influence of country of origin in the process of party identification acquisition". *Journal of Race, Ethnicity, and Politics* 9 (1): 80–98.

Jiménez, Tomás R, Deborah J Schildkraut, Yuen J Huo, and John F Dovidio. 2021. *States of belonging: Immigration policies, attitudes, and inclusion*. Russell Sage Foundation.

Kellstedt, Lyman A and James L Guth. 2021. "Religious voting in the 2020 presidential election: Testing alternative theories". *Politics and Religion Journal* 15 (2): 257–284.

Kerevel, Yann P. 2011. "The influence of spanish-language media on latino public opinion and group consciousness". *Social Science Quarterly* 92 (2): 509–534.

King, Gary, Robert O Keohane, and Sidney Verba. 2021. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Kolenikov, Stas. 2016, aug 1. "Post-stratification or non-response adjustment?". *Survey Practice* 9 (3).

Kuriwaki, Shiro, Stephen Ansolabehere, Angelo Dagonel, and Soichiro Yamauchi. 2024. "The geography of racially polarized voting: Calibrating surveys at the district level". *American Political Science Review* 118 (2): 922–939.

Laird, Ismail K and Chryl N White. 2021. *STEADFAST DEMOCRATS : how social forces shape black political behavior*. Princeton University Pres.

Lee, Lisa, Justine Bulgar-Medina, Kristen Neishi, Angela Houghton, and Manal Sidi. 2022. "Strategies for increasing asian american, native hawaiian, and pacific islander representation in survey research". *Survey Practice* 15 (1).

Lee, Sunghee, Hoang Anh Nguyen, and Jennifer Tsui. 2011. "Interview language: a proxy measure for acculturation among asian americans in a population-based survey". *Journal of immigrant and minority health* 13 : 244–252.

Masuoka, Natalie, Kumar Ramanathan, and Jane Junn. 2019. "New asian american voters: Political incorporation and participation in 2016". *Political Research Quarterly* 72 (4): 991–1003.

McDaniel, Jason. 2018. "Does more choice lead to reduced racially polarized voting? assessing the impact of ranked-choice voting in mayoral elections". *California journal of politics and policy* 10 (2).

McKenzie, Brian D and Stella M Rouse. 2013. "Shades of faith: Religious foundations of political attitudes among african americans, latinos, and whites". *American Journal of Political Science* 57 (1): 218–235.

Ngo-Metzger, Quyen, Sherrie H Kaplan, Dara H Sorkin, Brian R Clarridge, and Russell S Phillips. 2004. "Surveying minorities with limited-english proficiency: does data collection method affect data quality among asian americans?". *Medical Care* 42 (9): 893–900.

O'Hegarty, Michelle, Linda L Pederson, Stacy L Thorne, Ralph S Caraballo, Brian Evans, Leslie Athey, and Joseph McMichael. 2010. "Customizing survey instru-

ments and data collection to reach hispanic/latino adults in border communities in texas". *American Journal of Public Health* 100 (S1): S159–S164.

Park, David K, Andrew Gelman, and Joseph Bafumi. 2006. "State-level opinions from national surveys: Poststratification using multilevel logistic regression". *Public opinion in state politics*: 209–28.

Perl, Paul, Jennifer Z Greely, and Mark M Gray. 2006. "What proportion of adult hispanics are catholic? a review of survey data and methodology". *Journal for the Scientific Study of Religion* 45 (3): 419–436.

Perez, Efren. and Margit. Tavits. 2022. *Voicing Politics How Language Shapes Public Opinion.* Princeton Studies in Political Behavior Ser. ; v.45. Princeton: Princeton University Press.

Roman, Marcel F. 2023. "Living in the shadow of deportation: How immigration enforcement forestalls political assimilation". *Political Research Quarterly* 76 (3): 1460–1474.

Rothstein, Richard. 2017. *The color of law: A forgotten history of how our government segregated America.* Liveright Publishing.

Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih. 2023. "Cooperative Election Study Common Content, 2022".

Schaffner, Brian, Marissa Shih, Stephen Ansolabehere, and Jeremy Pope. 2025. "Cooperative Election Study Common Content, 2024".

Schildkraut, Deborah J. 2013. *Press" one" for English: language policy, public opinion, and American identity*. Princeton University Press.

Simpson, Andrea Y. 1998. *The tie that binds : identity and political attitudes in the post–civil rights generation*. New York: New York University Press.

Slaughter, Christine, Chaya Crowder, and Christina Greer. 2024. "Black women: Keepers of democracy, the democratic process, and the democratic party". *Politics & Gender* 20 (1): 162–181.

Stout, Christopher T and Jennifer R Garcia. 2015. "The big tent effect: descriptive candidates and black and latino political partisanship". *American Politics Research* 43 (2): 205–231.

Swope, Carolyn B, Diana Hernández, and Lara J Cushing. 2022. "The relationship of historical redlining with present-day neighborhood environmental and health outcomes: a scoping review and conceptual model". *Journal of Urban Health* 99 (6): 959–983.

Texas Legislature . 2023, 01. "Plan c2193". *Capitol Data Hub*.

Tokeshi, Matthew. 2023. "Anti-black prejudice in asian american public opinion". *Politics, Groups, and Identities* 11 (2): 366–389.

U.S. Census Bureau. 2023. "Public use microdata sample".

Watts, Donovan A. 2024. "Black millennials, slipping alliances, and the democratic party". *Journal of Race, Ethnicity, and Politics* 9 (2): 211–234.

Wilkinson, Betina Cutaia, James C Garand, and Johanna Dunaway. 2015. "Skin tone and individuals' perceptions of commonality and competition with other racial and ethnic groups". *Race and Social Problems* 7 : 181–197.

Wong, Janelle S, S Karthick Ramakrishnan, Taeku Lee, Jane Junn, and Janelle Wong. 2011. *Asian American political participation: Emerging constituents and their political identities*. Russell Sage Foundation.

Yeung, Yat To and Julian Wamble. 2024. "Are we one? the perception of language appeal targeting asian americans". *Working Paper, SSRN*.