PPIC PUBLIC POLICY INSTITUTE OF CALIFORNIA
30 YEARS

Do Registration Reforms Add New Voters or Keep Californians Registered?

Technical Appendices

CONTENTS

Eric McGhee, Jennifer Paluch, and Mindy Romero

Appendix A. Data Sources and Formatting

Merged California registration files

To identify movers and new registrants, we merge snapshots of the California voter file from multiple points in time to each other. A registrant's record after a move might contain differences from the original: either subtle ones due to error, or meaningful ones like a new last name or a nickname. Thus, an exact merge would be likely to miss a large number of records, so we turn to more sophisticated Bayesian "fuzzy" matching with the fastLink package for R (Enamorado, Fifield, and Imai. 2019). fastLink conducts approximate matches and makes it easy for the user to block the data into more compatible subsets to speed the matching process. The blocking is essential to make the problem tractable given the size of the California voter file.

Prior to conducting the match, we also run the wru package for R (Imai and Khanna 2016) to impute race for each registrant based on last name, gender, location, and party registration. fastLink also allows the user to set a closeness parameter for the fuzzy match and to designate variables for full or partial matches. Since we expected the most errors in the first and last name fields, we set those to be partial matches, and used full matches for middle name, birth date, and voting history (which is also a one-character string). Based on experiments with different settings, we set conservative values for the closeness parameters: 0.94 for partial match and 0.99 for full match. This likely cost us some matches we would otherwise have found, but our examination of samples from lower values suggested far more false positives than we were comfortable accepting.

With these data and this tool, we follow a multi-step process to ensure we catch as many matches as possible:

- *Unchanged records:* Exact match by last name, first name, middle name, birth date, and latitude/longitude
- *Simple match for movers statewide:* Exact match by last name, first name, middle name, and birth date
- *Deeper search for within-county movers:* Fuzzy match by last name, first name, middle name, birth date, and recent voting history, separately by county and gender
- *Deeper search for cross-county movers:* Fuzzy match by last name, first name, middle name, birth date, and recent voting history, separately by gender
- *Final sweep for movers:* Fuzzy match by last name, first name, middle name, and birth date, separately by gender
- *Final sweep for recently married women who changed their surname:* Exact match by first name, birth date, and latitude/longitude, separately by gender

The matches beyond the first two exact matches added 1.2 million more matches between 2012 and 2016, and 2 million between 2016 and 2020.

The 2012-2016 merge serves as a calibration for the kind of change in the file we would normally expect from our merge in the absence of California's registration policy changes. This is important because new registrants and those that drop off the file might just be movers from in or out of state and not truly "new" or "dropped off" at all. It will also serve to calibrate for any bias in the merging process, so long as we assume such bias is approximately the same before and after the adoption of AVR and so can be differenced out. There certainly remains some unavoidable random error to the merge that we cannot fully difference out, but it should have the effect of attenuating our estimates and making our conclusions more conservative.

Population denominators

Our nationwide registration data come from the data vendor Catalist. Catalist maintains a large national file that combines all state-level files together. They add imputations for race and ethnicity to this file. They also merge this file to itself over time, allowing them to see who has moved, stayed put, or is entirely new to voter registration (meaning they have never appeared before in any county file anywhere in the United States). We obtained county-level aggregates of snapshots from the 2012, 2016, and 2020 presidential general elections. These aggregates included new registrants and cross-state, cross-county, and within-county movers, all since the previous presidential election in each case.

For our "adjusted" estimates, we needed denominators that reflected the total population potentially affected by the registration policy changes. For address updates, this denominator would consist of all registered voters who had moved within the state since the previous presidential election. For new registrations, it would consist of all eligible residents who had not been registered in each county four years previously, either because they were in the county but not eligible, in the county and eligible but not registered, or eligible but not in the county.

In our nationwide county-level analysis, the denominator of movers started with a rough imputation combined from a variety of sources. We used the 2008-2012, 2012-2016, and 2016-2020 county-to-county migration tables from the U.S. Census Bureau to obtain the average number of people who had moved into or out of each county (both to other states and other counties) in the past year for each of those overlapping 5-year periods. To convert these total population numbers to the citizen voting age population (CVAP) relevant to our analysis, we calculated CVAP rates for each type of mover from the corresponding IPUMS file (Ruggles, et al. 2022) and multiplied these rates times the numbers from the migration tables. The IPUMS samples include only large counties individually and aggregate all others, so for the aggregated counties we applied the CVAP rate for the entire aggregated area to each constituent county individually.

Only migrants moving within the state (county-to-county or within-county) were relevant to our address update analysis. All other migrants were either ineligible for address updates from the state's AVR system because they left the state entirely or were more properly considered newly-eligible residents because they had just moved into the state. Moreover, address updates apply only to migrants who were already registered; to reflect this population, we multiplied our within-state migration totals by the registration rate at the same point in time (based on the CVAP and David Leip registration sources described below).

This process provided estimates of CVAP and registration rates for migrants in each county, but only by applying rates for higher levels of aggregation to individual counties. We turned to iterative proportional fitting (Lomax and Norman 2016) to develop more accurate intersections of registration and CVAP by migration rate, and to control these numbers to the county totals. The estimates from the process above served as our "sample" weights, and we fit those iteratively to Census county migration aggregates and CVAP-registration breakdowns from the Census and David Leip. To make the fitting process more tractable, we collapsed the migration groups into movers and non-movers, and the CVAP-registration categories into ineligible, eligible and unregistered, and registered (i.e., combining children with non-citizens into a total ineligible category). The implied CVAP and registration rates among movers were then applied to all categories of movers in a given county. We used these numbers as our estimates of the unregistered population for all counties, and as our estimates of the eligible population for counties that were not individually represented in IPUMS.

Finally, for eligibility the product of these calculations represented the number of migrants for an average year in the 5-year aggregation period of the sample, so we multiplied each of our estimates by four to get the approximate number of total migrants in the four years between each set of presidential elections.

For estimates of eligible but not previously registered residents in our nationwide county-level analysis, we started by calculating the eligible-but-unregistered population in the county at the time of the previous presidential election, separately for each election year (i.e., the difference between total CVAP and total registered voters). The CVAP estimates for these calculations came from the Census Bureau's special CVAP tabulation of each entire 5-year ACS sample. The registration numbers came from David Leip's Atlas of Presidential Elections. We used Leip's data to extend our time series back to 2008, so we conduct the calculation for the 2012 election. However, this also means we were unable to generate these denominators separately for race, ethnicity, and age.

Next we "aged" this lagged estimate of eligible-but-unregistered by adding in newly eligible residents from the ensuing four-year period and subtracting eligible-but-unregistered people who moved away or died. The newly-eligible residents come from three sources: 18-21 year-old citizens who came of voting age in the four years since the previous presidential election; eligible migrants into the county; and recently naturalized immigrants:

- Total counts of 18-21 year-olds come from the Census Bureau's tables of the entire ACS sample. We first multiplied these counts by CVAP rates from IPUMS as a starting point, then controlled these numbers to the age and CVAP county totals using iterative proportional fitting.
- The eligible migrants moving into the county also come from the migration data process described above. However, in contrast to address updates we multiply those totals by the share of eligible residents who were *unregistered*, since our analysis of new registrants is focused on those who are not already registered to vote.
- For recently naturalized immigrants, the Census Bureau's complete county-level tables only contain naturalizations grouped by fixed 5-year increments (e.g., 2005-2009, 2010-2014, 2015-2019), while IPUMS has naturalizations in the previous year but only for large counties and county aggregates. To transfer previous-year naturalization information to individual counties, we aggregated the complete tables to the same aggregation units as IPUMS and merged the two files. We then regressed logged totals for single-year naturalizations on logged total naturalizations from the previous two five-year aggregation periods (e.g., for the 2016-2020 file, we used the 2015-2019 and 2010-2014 aggregation periods; for the 2008-2012 file we used the 2010-2014 and 2005-2009 aggregation periods). The model interacted everything by the year of the IPUMS file. We then used the results of this regression to impute single-year naturalization numbers for the individual counties from the complete Census Bureau tables.

On the other side of the ledger are those who moved out or died:

- For migrants out of the county, we applied the same process as for migrants coming into the county, including a subset to those who are unregistered but eligible.
- For deaths we used county-level data from the Centers for Disease Control and Prevention (CDC). Deaths came grouped into five-year bins. These bins combine 18 and 19 year-olds, who are old enough to register, with 15-17 year-olds, who are not. However, deaths are extremely uncommon among late teens: the crude death rate for 15-19 year-olds is just 0.05%, and their share of all deaths in the data is just 0.23%. Thus for our purposes, 20 is a reasonable cutoff for adulthood that only slightly understates the adjustment. We multiplied these death totals by the county registration rate to obtain deaths among registered and unregistered.

Because this county imputation process is complex and includes multiple approximations layered on top of each other, we bootstrapped standard errors for models using these denominators.

In our California voter file analysis, the denominator of movers started with 1-year ACS data from IPUMS (Ruggles, et al. 2022). These data allowed us to identify total population, citizen voting age population (CVAP), movers into and out of the state, movers within the state, and immigrants naturalized in the previous year. We aggregated these groups by year, birth year, and race/ethnicity. For movers and naturalized immigrants, we aggregated the totals for each of the four years between presidential elections for an estimate of the total migration

and naturalization that had occurred. For total population and CVAP, we used the number at the point in time of the second presidential election in each comparison (2016 for 2012 and 2016; 2020 for 2016 and 2020). With CVAP by single year of age we were also able to identify the total number of 18 to 21-year-olds who had aged into voting eligibility by the point of the second presidential election in each pair. Death counts once again came from the CDC, except we downloaded single year of age instead of age bins.

National analysis

To confirm that broader shifts in the national political environment are not driving the California results, we use the data from Catalist for a two-way fixed-effects (TWFE) model. This model includes dummies for counties and election years to account for fixed differences between counties and uniform change across all counties over time. We also control for two additional registration reforms that might help explain any changes in registration patterns to the extent that they were adopted at the same time as AVR: 1) election day registration, where voters can register at the polls on election day; and 2) online voter registration, which makes it easier to sign up or update registration by offering a single online portal for the process. Finally, we control for the two-party statewide presidential vote margin to capture swing state mobilization effects. Full results for this specification with each of the outcome variables can be found in Appendix B.

Beginning with the 2020 election, Utah enacted a data privacy law that now limits the records that can be shared with outside parties, including voter registration records and vote history information. See https://vote.utah.gov/voter-privacy-information/ for details. This policy change made Utah's 2020 totals in the Catalist data migration data incomparable with previous years. The Catalist migration data suggested a drop of about 200,000 registrants between 2016 and 2020, while the state itself reported an *increase* of about 200,000 (or around 8%) over the same period of time. Thus, we drop this state from our analysis.

Appendix B. Regression Results

TABLE B1

Model results—all U.S. counties in presidential elections, 2012-2020

	Registration rate	Address updates: CVAP	New registrations: CVAP	Address updates: Potential	New registrations: Potential
Intercept	0.876	0.160	0.133	0.408	0.530
	(0.032)	(0.018)	(0.015)	(0.057)	(0.090)
AVR	0.024	0.010	0.012	0.052	0.035
	(0.003)	(0.004)	(0.001)	(0.011)	(0.008)
Election day registration	0.019	0.012	0.006	-0.006	0.022
	(0.004)	(0.004)	(0.002)	(0.013)	(0.009)
Online voter registration	0.005	-0.005	0.003	-0.008	-0.006
	(0.002)	(0.002)	(0.001)	(0.006)	(0.005)
Presidential vote: statewide margin	-0.218	-0.001	-0.076	-0.097	0.026
	(0.017)	(0.021)	(0.006)	(0.063)	(0.045)
County fixed effects	X	X	X	X	X
Year fixed effects	X	X	X	X	X
RMSE	0.043	0.048	0.016	0.131	0.099
N	8977	8977	8977	8974	8744

SOURCES: Catalist (registration migration data); David Leip's Atlas of U.S. Presidential Elections (lagged registration for denominator); U.S. Census Bureau (CVAP, migration, and naturalizations for denominator); IPUMS (CVAP for migration and naturalization groups in the denominator); Center for Disease Control and Prevention (death counts for denominator); Federal Election Assistance Commission (registration policies); National Conference of State Legislatures (registration policies).

NOTES: Cell entries are ordinary least squares coefficients with standard errors in parentheses. The outcome variables in each column are as follows: "Registration rate" = total registration as a share of total CVAP; "Address updates: CVAP" = address updates as a share of total CVAP; "New registrations: CVAP" = new registrations as a share of total CVAP; "Address updates: Potential" = address updates as a share of movers; "New registrations: Potential" = new registrations as a share of eligible Californians who were not registered in California four years earlier. Full descriptions of these denominators can be found in Appendix A. For the registration rate and the CVAP denominators, standard errors are clustered by state. For the denominators that reflect the potentially affected community, the standard errors are bootstrapped with 1000 random draws, to better reflect the full uncertainty of the imputation process described in Appendix A.

TABLE B2

Imputation of previous year naturalizations from binned past naturalizations

	Outcome: Logged naturalizations in previous year
Intercept	-1.45
	(0.211)
Log naturalizations, lag period 1	0.767
	(0.076)
Log naturalizations, lag period 2	0.243
	(0.078)
Year = 2016	-1.104
	(0.281)
Year = 2020	-1.161
	(0.275)
Log naturalizations, lag 1 period X Year = 2016	0.356
	(0.151)
Log naturalizations, lag 1 period X Year = 2020	0.064
	(0.130)
Log naturalizations, lag 2 period X Year = 2016	-0.352
	(0.151)
Log naturalizations, lag 2 period X Year = 2020	-0.034
	(0.133)
RMSE	0.695
N	1331

SOURCES: U.S. Census Bureau (grouped period naturalizations); IPUMS (single previous year naturalizations)

NOTES: Cell entries are ordinary least squares coefficients with standard errors in parentheses. The first lag period is 2015 or later for the 2020 data, and 2010 or later for the 2012 and 2016 data. The second lag period is 2010-2014 for the 2020 data, and 2005-2009 for the 2012 and 2016 data

PPIC

PUBLIC POLICY
INSTITUTE OF CALIFORNIA

30 YEARS

The Public Policy Institute of California is dedicated to informing and improving public policy in California through independent, objective, nonpartisan research.

Public Policy Institute of California
500 Washington Street, Suite 600
San Francisco, CA 94111
T: 415.291.4400
F: 415.291.4401
PPIC.ORG

PPIC Sacramento Center
Senator Office Building
1121 L Street, Suite 801
Sacramento, CA 95814
T: 916.440.1120
F: 916.440.1121