# Using Nationwide Voter Files to Study the Effects of Election Laws*

Bernard L. Fraga† John B. Holbein‡ Christopher Skovron§

July 17, 2018

[*This is a very preliminary draft; please do not circulate or cite.*]

**Abstract**

Voter file-derived datasets have become a common source of information used by researchers studying voting behavior. Despite the various advantages offered by such data, however, survey-based analyses continue to dominate the study of election laws. While concerns about cost and availability are paramount for researchers considering the use of voter file data, less attention has been paid to the methodological advantages and challenges of using voter lists for elections research. In this paper, we outline the various methodological considerations encountered when using voter file data to estimate the causal effect of state-level election laws on voter turnout. We do this while describing the features of a national, comprehensive voter file compiled by the Data Trust, a data vendor. We provide information about the quality of the dataset and its accompanied measures, both modeled and from administrative sources, through accessing the individual-level data on over 200 million registered voters.

*Prepared for presentation at the 2018 Election Sciences, Reform & Administration Conference. We wish to thank the Data Trust for providing access the voter file data used in this study. We also wish to thank Brittany Bair, Drew Brown, Matthew Baldwin, Alessandra Carbajal, Amanda Eatough, and Nichole Keddington for their outstanding work as research assistants. Holbein wishes to thank the National Science Foundation for their generous financial support (SES-141681 and SES-1657821).

†Assistant Professor, Indiana University. bfraga@indiana.edu.

‡Assistant Professor, Brigham Young University. john_holbein@byu.edu.

§Postdoctoral Scholar, Northwestern University, cskovron@northwestern.edu.

# Introduction

For decades, the standard approach to studying the effects of state-level election laws, policies, and reforms has been to examine differences in voter turnout across states as measured by large-scale surveys. From Wolfinger and Rosenstone's (1980) analysis of registration laws using the 1972 and 1974 Current Population Survey November Supplements, to Hajnal, Lajevardi and Nielson's (2017) research on the effect of voter identification laws using 2006-2014 Cooperative Congressional Election Study data, survey-based studies of election laws have followed in the tradition of American political behavior research that continues to inform our understanding of our political activities (Campbell et al. 1960; Verba, Schlozman and Brady 1995; Rosenstone and Hansen 1993; Leighley and Nagler 2013).

Yet while surveys continue to dominate the study of election laws, the broader study of turnout is in the midst of a transition to voter registration list or "voter file"-based analyses (Sigelman et al. 1985; Gimpel, Dyck and Shaw 2004; Dyck and Gimpel 2005; Green and Gerber 2005; McDonald 2007; Cooper, Haspel and Knotts 2009; Ansolabehere and Hersh 2012; Hersh 2015; Fraga 2016*a*,*b*; Holbein and Hillygus 2016). Over the past 10 years, approximately half of the U.S. voter turnout-focused studies in the *American Political Science Review*, *American Journal of Political Science*, and the *Journal of Politics* made use of voter file-based data (Fraga, Spahn and Yan 2017); over three dozen articles in all. There are many reasons for this transition, including, but not limited to, issues with using self-reported rather than validated voting, the difficulty of examining subgroups of interest even with large surveys, and the trend toward design-based inference (including experimental methods) when studying voter behavior. Voter file-based analyses may be particularly appropriate for the study of election laws and turnout, as state or county-level treatments further necessitate attention to these issues (Keele and Minozzi 2013; Highton 2017; Burden 2018).

While the solution to many of these problems may lie in compiling comprehensive individual-level voter records, such an achievement has been out of reach for most researchers. Voter-file based studies are often restricted to the one or few states where voter file data is available and not prohibitively expensive (McDonald 2007; Cooper, Haspel and Knotts 2009; McDonald 2017*a*), or relies on third-party vendors who generally provide samples from their database or aggregate counts at a higher level than the individual voter (Hersh 2015). However, as each year passes, it seems that the number of researchers with access to a complete national file grows.[1] Of course, nationwide individual-level voter files are no panacea—coming with their own issues and difficulties (Nyhan, Skovron and Titiunik 2017). Simply acquiring a nationwide voter file is not enough, as using one to estimate the effects of election law comes with many potential challenges. As we will argue below, "big data" is no solution (on its own) to problems involving causal identification.

In this paper we address the data and design issues that underlie the estimation of election law effects with nationwide voter files. We talk about the challenging coding and data issues that one must face when using these large datasets. We also discuss the new designs—beyond a simple difference-in-difference—that can be brought to bear that bring an additional layer of causal robustness to the study of electoral reforms, including: individual fixed effects, individual-level pairwise matching, and synthetic controls. We discuss these methods' underlying strengths, weaknesses, and assumptions. We do this in the context of a discussion of a nationally comprehensive list of individual-level registered voters compiled by the Data Trust. As we discuss below, we have obtained a copy of this list and are currently exploring its promise for future work on electoral rules specifically, and voter turnout more broadly. Below we describe the quality of the dataset and its accompanied measures, both modeled and from administrative sources. At present, we are looking for feedback on additional checks that we can and need to run to explore the validity/usefulness/strengths/weaknesses of this data source.

We conclude by arguing that though there are challenges to working with a nationwide voter file, the benefits outweigh the costs. No individual-level dataset comes without unique challenges. Because it is unlikely to impossible to conduct randomized control trials varying citizens' exposure to election laws, no identification strategy will come without assumptions/weaknesses in estimat-

[1]We know of two other research teams (one at Princeton and one at Harvard) that have a complete individual-level nationwide voter file, along with one other effort to compile such a dataset for academic research.

ing the causal effects of electoral laws. However, nationwide voter files paired with the methods we suggest here offer great value in pushing discussions of the effects or non-effects of electoral rules forward.

## Limitations of Survey Data in Estimating Election Law Effects

As noted above, much research on voter turnout has transitioned away from (sole) reliance on survey data. Below we discuss some of the key issues facing survey-based analyses, with an emphasis on how these issues may impact our understanding of the turnout effect of state-level election laws.

The clearest (and perhaps most researched) issue with using surveys to measure the effect of election laws is the self-reported nature of the dependent variable. The misreporting (generally overreporting) of voter turnout is broadly understood to have a substantial impact on survey-based estimates of both how many individuals participate (Burden 2000; McDonald 2007), and what factors predict participation (Silver, Anderson and Abramson 1986; Bernstein, Chadha and Montjoy 2001; Ansolabehere and Hersh 2012, but see Berent, Krosnick and Lupia 2016). Attempts to validate self-reported measures through matching respondents with voter file data are often seen as one way of dealing with this issue, as in Hajnal, Lajevardi and Nielson (2017). However, even when high-quality surveys like the American National Election Studies or Cooperative Congressional Election Studies are matched to voter records, issues remain (Burden 2018). Taking this approach introduces new problems involving what to do with individuals who are not matched with voter records and how far one can use these samples to draw inferences about state-level differences or subgroup differences of interest (Grimmer et al. 2018). The CPS November Supplement, which may be able to overcome some of these sampling issues, does not have validated voting.[2]

Even if we are able to limit problems related to self-reporting, survey-based analyses of voter

[2]Hur and Achen (2013) propose adjusting CPS estimates to account for observed state-level rates of turnout. However, this adjustment assumes that rates of misreporting are equivalent across groups within a state, a claim that is difficult to assert given the extant literature on demographic differences in who misreports voting (Silver, Anderson and Abramson 1986; Ansolabehere and Hersh 2012; McKee, Hood and Hill 2012). Moreover, the CPS has other issues with representativeness at geographies lower than the state level.

turnout face another hurdle when seeking to establish election laws as *producing* measurable effects. In general, it is difficult to establish causal relationships with observational data (Morgan and Winship 2007). However, these issues are all the more obvious when seeking to compare behaviors across states (Erikson and Minnite 2009; Keele and Minozzi 2013; Highton 2017; Burden 2018). Election laws, rules, and procedures do not change in a vacuum, meaning that regression-based inferences about the effect of election laws necessitates strong assumptions about our ability to adequately capture all of the factors that impact turnout at both the individual and state level (Burden et al. 2014; McDonald, Shino and Smith 2015; Burden et al. 2017). Keele and Minozzi (2013) suggest that this is virtually impossible, and that instead analysts should leverage over time variation in turnout *within* a state as a first step toward establishing causal relationships.[3] The canonical difference-in-differences approach (Morgan and Winship 2007), relying on the assumption that, e.g., two states would follow the same pattern of turnout over time in the absence of the implementation of an election law change, allows for much stronger inferences (Highton 2017). Again, surveys fall short in this regard, as even large scale, multiyear survey data will contain enough variation in sampling to make it difficult to track small changes in turnout across states (Grimmer et al. 2018). Moreover, if one wants to leverage stronger research designs that utilize variation along other dimensions—such as at the individual-level or close to geographic discontinuities—surveys may not be ideal given their (frequently) cross-sectional design representative only at the state level (at the lowest level of geography).

Alternatives to survey data may address at least some of these issues. Research spanning the last decade does indeed indicate that voter file data, including voter file-derived data curated by data vendors, provides a far more accurate assessment of how many people vote than surveys (Burden 2000; McDonald 2007; Fraga, Spahn and Yan 2017). States generally provide linked turnout history of individual registrants as well, allowing for over time analyses (Cooper, Haspel and Knotts

[3]One response may be that researchers are not seeking to establish causal relationships, instead only producing correlations suggestive of a pattern that might inform our understanding. This is not a panacea, as we have an extensive, separate literature on what predicts various election laws from being implemented (Biggers and Hanmer 2017; Hicks et al. 2015; Hicks, McKee and Smith 2016) such that a correlational approach not only fails to establish a causal link, but may *imply* spurious relationships that distract from the matter at hand.

2009). On these marks, voter file-derived data has a clear advantage over surveys.[4] That said, voter files as provided by states do not provide the same suite of individual-level demographic variables provided by surveys (Cooper, Haspel and Knotts 2009; Hersh 2015). While the imputation of individual-level traits from geographic data is one option (Hersh and Nall 2016), ecological fallacies may challenge the validity of this approach (Imai and Khanna 2016). To conduct the subgroup analyses that are an important focus of previous work (Erikson and Minnite 2009; Dropp 2013; Hajnal, Lajevardi and Nielson 2017), modeling procedures leveraging individual-level detail may be necessary.

Given these concerns, we believe that *individual-level voter file data* provides a tremendous opportunity for researchers seeking to understand the effect of election laws to conduct analyses with greater internal validity than possible with survey data. Furthermore, a *national* database may enable external validity approaching, and perhaps exceeding, that provided by surveys such as the CPS—a critical point when examining state-level effects since there are only 50 cases in an election year. Again, we are not the first researchers to note the potential for voter file data to improve our understanding of electoral behavior. Beyond the 10 state analysis of raw voter file data conducted by McDonald (2007), Ansolabehere and Hersh (2014) and Hersh (2015) examine features of the 50-state database compiled by Catalist, LLC, a leading voter file vendor for academic research. Our goal here is to provide a similarly executed check of the quality of voter registration lists compiled by a source (Republicans) that has received much less attention in the literature. We thus begin by conducting similar tests with the Data Trust national voter file before outlining recommendations for how such data should be used to examine the effects of election laws on turnout. We then turn our attention to methodological opportunities for using nationwide voter files in the estimation of electional laws.

[4] Of course, the analysis of voter turnout using state or county-provided counts of ballots cast is a powerful alternative (McDonald and Popkin 2001; McDonald 2017*b*), though turnover in state populations (including deaths and relocation across states) again introduces challenges when seeking to derive causal effects with over time data.

# Data: The Data Trust National Voter File

In order to use nationwide voter files to estimate the effects of election laws, it is necessary that the voter files be of high quality. It's entirely possible that voter lists, if they are not comprehensive or have other problems with measurement, may introduce bias or error in estimates of the effect of election laws. In this section, we pay specific attention to potential threats of this variety. We start by first assessing overall registration counts, we then look at demographic comparisons, explore missingness in variables of potential interest, examine discrepancies in turnout numbers with official counts, and benchmark our numbers to other known quantities. We address specific potential challenges and assess their potential to introduce bias.

To explore some of the issues that might influence election law estimates, we use newly available nationwide voter file maintained by The Data Trust, a private vendor that contracts with Republican clients. The Data Trust collects and cleans data on all registered voters in the United States, as well as many other adults. We have access to a complete national file that reflects data updates following the 2016 election (in most states) and was compiled in September 2017. Unlike some other academic analyses of voter file data which often rely on 1% samples or specific queries to other quantities of interest (e.g., Hersh 2015), we have access to the entire individual-level national voter file collected by Data Trust (N $\approx$ 200 M). At present, we only have access to a single, nationwide snapshot; however, we have an agreement in place with the Data Trust to get future and historical snapshots annually. For this reason, all the analysis below come from the 2017 build.

## Registration Counts

A first basic chest for the quality of voter file data is whether the voter file is missing large swaths of the electorate. If certain states or demographics were to be systematically missing from the file, the missingess will introduce bias into our estimates and limit our ability to generalize to populations of interest.

Table 1 begins this process for the Data Trust voter file by showing the breakdown of individuals across several relevant characteristics. As can be seen, most of the Data Trust's sample comes from voter registration files. About 99% of the 201 million individuals in the Data Trust file are registered to vote. While there are some individuals merged in from vendor files who are not registered, the overall number of these individuals is small. It would seem, then, that the Data Trust relies mostly on voter registration files to build its sample, and less on vendor data. Also of note in Table 1 is the number of individuals that the Data Trust says are registered to vote. The Data Trust lists just under 199 million registered voters. This count is very close to that provided by other voter file vendors: as a reference, TargetSmart (another nationwide voter file vendor) recently publicly reported just over 200 million registered voters.[5] By this metric, the Data Trust appears to perform well.

Figure 5 plots the number of registered voters available in each state. The graph is sorted by the number of citizens registered to vote (dark bars), with the number of voting age population estimates also shown as a benchmark. As can be seen, the number of registered voters at the state level correlates strongly with estimates of the number of people of voting age in the state as a whole ($R = 0.98$). This strong correlation with the voting age population suggests that we probably aren't missing large swaths of the registered population across states.[6]

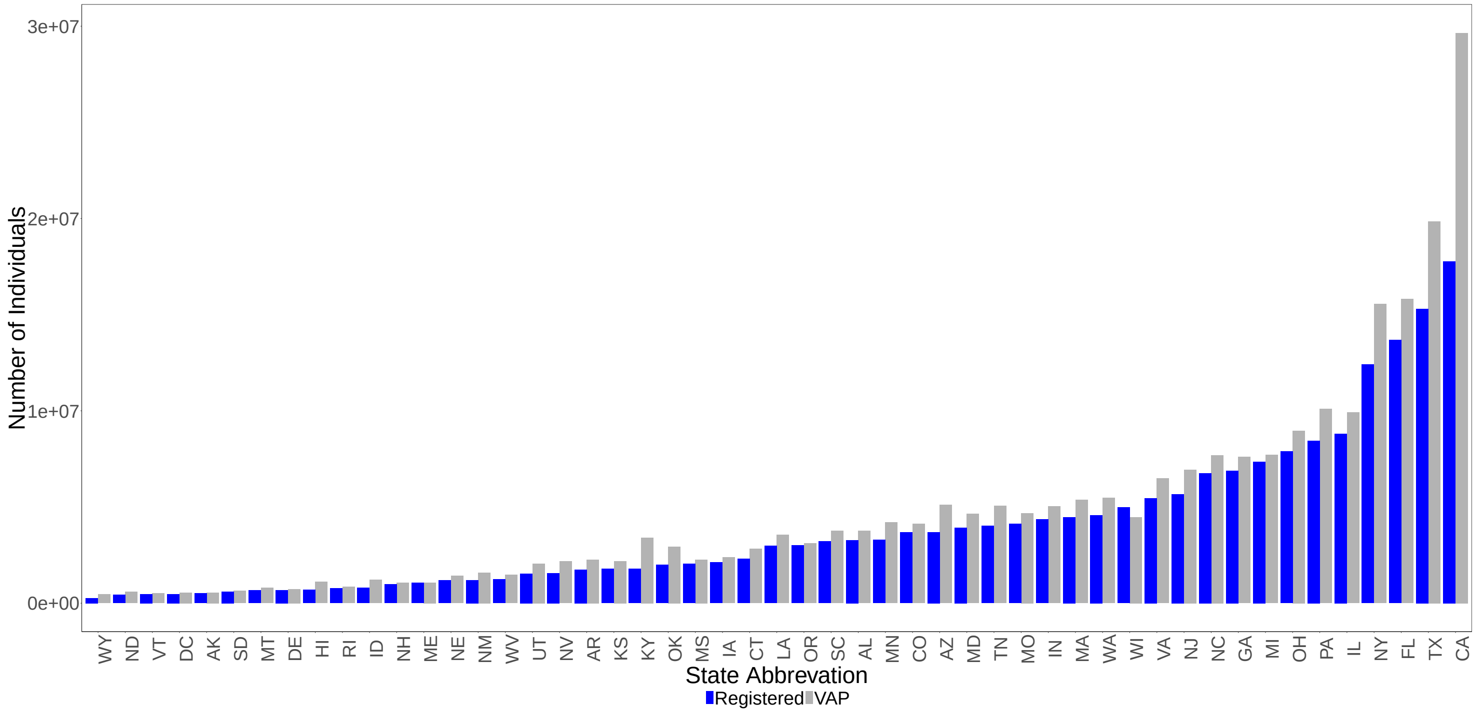
[5]See "America hits new landmark: 200 million registered voters" *Politico*, 10/19/2016.

[6]Registration doesn't necessarily have to go with population size. Some states with similar population sizes (e.g. NY and FL) may have vastly different registration rates, for a variety of reasons. (Registration rates can be seen by looking at the gap between the grey bars and the blue bars within a given state.) However, the fact that we, on average, see more registrants where there is a larger pool of potential voters is comforting.

Table 1: Overview of Quantities of Interest in the Data Trust Files

| Field | Count | Percent of File |
|---|---|---|
| **Totals** | | |
| Total in File | 201,007,802 | 100% |
| **Registration Status** | | |
| Number Registered | 198,929,364 | 99.0% |
| Number Not Registered | 2,078,438 | 1% |
| **Political Party** | | |
| Registered/Modeled Republican | 58,012,736 | 28.9% |
| Registered/Modeled Democrat | 68,123,575 | 33.9% |
| Registered/Modeled Independent/Other | 74,871,491 | 37.3% |
| **Race** | | |
| Registered/Modeled Black | 22,867,599 | 11.4% |
| Registered/Modeled White | 126,566,122 | 63.0% |
| Registered/Modeled Hispanic | 19,213,356 | 9.6% |
| Registered/Modeled Other Race | 25,976,640 | 12.9% |
| Registered/Modeled Asian | 6,384,085 | 3.2% |
| **Age** | | |
| 18-29 | 31,799,371 | 15.8% |
| 30-40 | 33,999,962 | 16.9% |
| 41-50 | 29,565,266 | 14.7% |
| 51-60 | 33,511,787 | 16.7% |
| 61-70 | 29,059,635 | 14.5% |
| 71+ | 43,071,781 | 21.4% |
| **Data Checks** | | |
| Number of Individuals with Duplicate Name/DOB | 6,565,810 | 3.7% |
| Missing Geocode | 395,737 | 0.2% |
| Missing Birth Year | 3,325,632 | 1.7% |
| Missing Birth Month | 24,813,930 | 12.3% |
| Missing Birth Day of Month | 24,823,604 | 12.3% |
| Missing Registration Date | 24,627,308 | 12.2% |

Figure 1: Data Trust Registration Counts and VAP by State



## Demographic Comparisons: Race

Another issue that we could face with the nationwide voter file data in estimating electoral effects is whether the Data Trust is able to appropriately identify subgroups of interest. We address two of these here; first exploring race, then political party.

Despite the advantage in power that we gain from using the voter file relative to survey data, we face an important limitation in that most states do not record registrants' race/ethnicity. The Data Trust codes race in the following way: first, by incorporating individuals' self-reported race in the states where it is available. Then, to supplement this, they model race using individuals names and geographic locations paired with data from the Census.[7]

When it comes to racial estimates, Table 1 shows that the Data trust puts the percent of White citizens at around 63% and other racial groups comprising the rest of the population. In its estimate closest to the Data Trust snapshot date, the Census estimated that among the entire population 60.7% of is white.[8] (For reference, the respective census numbers for Blacks is 13.1% vs. 11,4% in the Data Trust file.) While this comparison is not perfect, it suggests that the topline estimates are close to one another. (Part of the reason for having a slightly higher percentage of whites in the file may be due to racial patterns in registration.)

Digging into this further, as can be seen in Figure 2, this approach to coding race benchmarks well with official reports at the state level. While not perfect (as would be case if all points where on the 45 degree reference line, the relationship is still quite strong regardless of whether we look at the state or county level (State $R = 0.94$; County $R = 0.97$.)

[7] In this draft of the paper, we rely on the Data Trust's coding of race. However, given the lack of transparency in the modeling process, we will eventually conduct our own race modeling. This will be done by using the `wru` package in `R` (Khanna, Imai and Jin N.d.). This approach assigns voters in these states to racial/ethnic groups based on name matching combined with geographic contextual information. We will identify voters as likely African American, Hispanic/Latinx, non-Hispanic White, Asian, or other. We will assign each individual to the group corresponding to the plurality probability, placing individuals who could not be predicted into the "other" category.

[8] See "Census Quick Facts," *United States Census Bureau*, 7/1/2017

Figure 2: Data Trust Black Variable and Census Black Population Counts

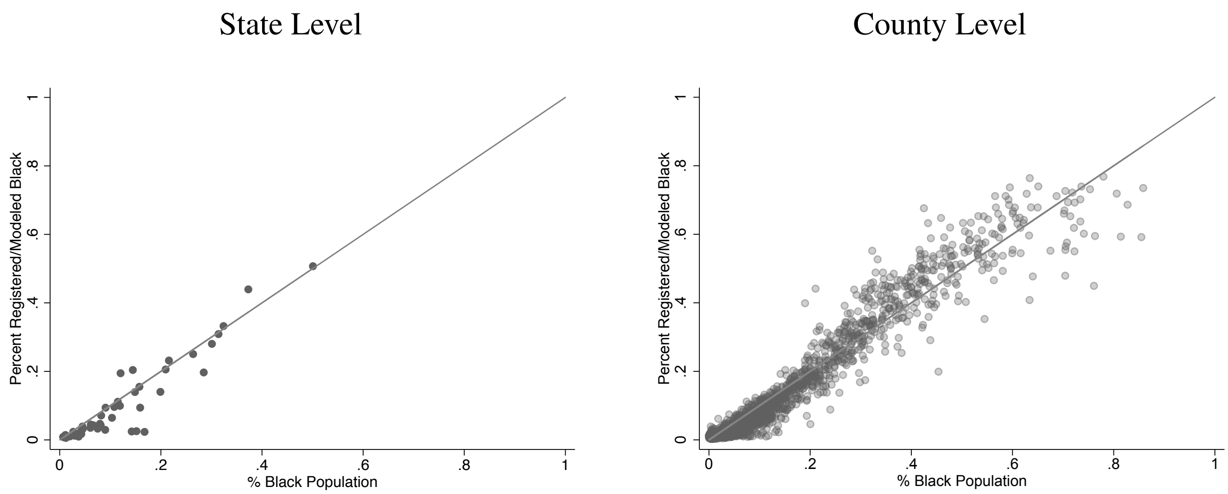



Figure 2 plots the relationship between the percent of the population that is African American (from the Census) and the Data Trust's race variable, which is a hybrid between registered race (where it is available) and modeled raced based on name and geography (Imai and Khanna 2016). It plots the relationship at two levels: state (left) and county (right). Pearson's R: 0.94 (left), 0.97 (right).

## Demographic Comparisons: Political Party

Consumer voter files allow us to look at one's political party several different ways: first, by looking only in the 33 states where party registration is available and, second, by looking at all states using the Data Trust's modeled party estimates. The Data Trust, like other vendors, have a strong incentive to get modeled party as close as possible to reality, as voter contact programs depend on getting party right. They do so by first incorporating party registration (when that is available), then by incorporating votes in primary elections, and then finally by culling and coding based on targeted voter contacts. As we are unlikely to leverage the information from the last piece and, as such, would inevitably do worse than the Data Trust at modeling party, we think using their modeled scores in conjunction with party registration alone is reasonable.

Using these methods, the Data Trust estimates that about 29% of individuals in its sample are Republicans, 34% are Democrats, and 37% are Independents/Others. The estimates are fairly accurate compared to survey data; for reference, Gallup's running series estimated in September

2017 (the same month as the Data Trust snapshot) that among the entire population those numbers are 29%, 30%, and 40% respectively.[9] The Data Trust seems to get very close to the Republican number and slightly overestimates the number of Democrats relative to Independents in the population. (Some of these differences might simply be because citizens who register are different from the population as a whole that Gallup is trying to measure/approximate)

In addition, two other factors suggest that modeled party is of high quality. First, Igielnik et al. (2018) link Pew survey data to commercial voter file vendors and show that party estimates at the individual-level are quite accurate, even in states where party registration is not available in the files. Second, we can benchmark the Data Trust's party measures/estimates with presidential vote returns at the state level. As can be seen in Figure 3, as best we can tell, the Data Trust's modeled party estimates stack up well with official returns. The correlation between modeled party and presidential vote share and modeled party is 0.82. The Data Trust appears to slightly underestimate the proportion of people voting Republican in the file; however, this appears to be fairly systematic and uniform across states.

[9] See "Party Affiliation," *Gallup* September 6-10 measurement.

Figure 3: Data Trust Modeled Party Relative to Presidential Vote Share (2016)

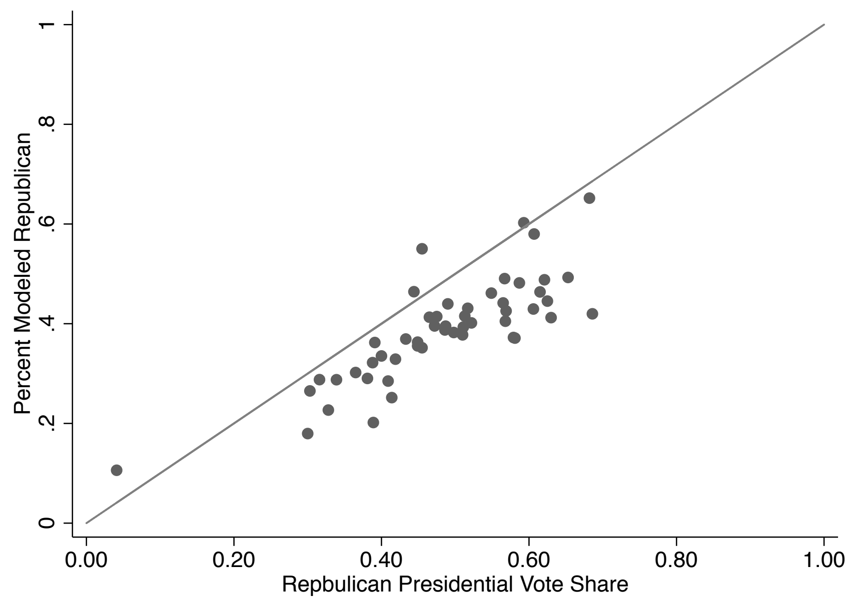



Figure 3 plots the relationship between Republican vote share (from official counts) and the Data Trust's modeled party score. Pearson's R: 0.82.

## Missingnesss in Variables of Import

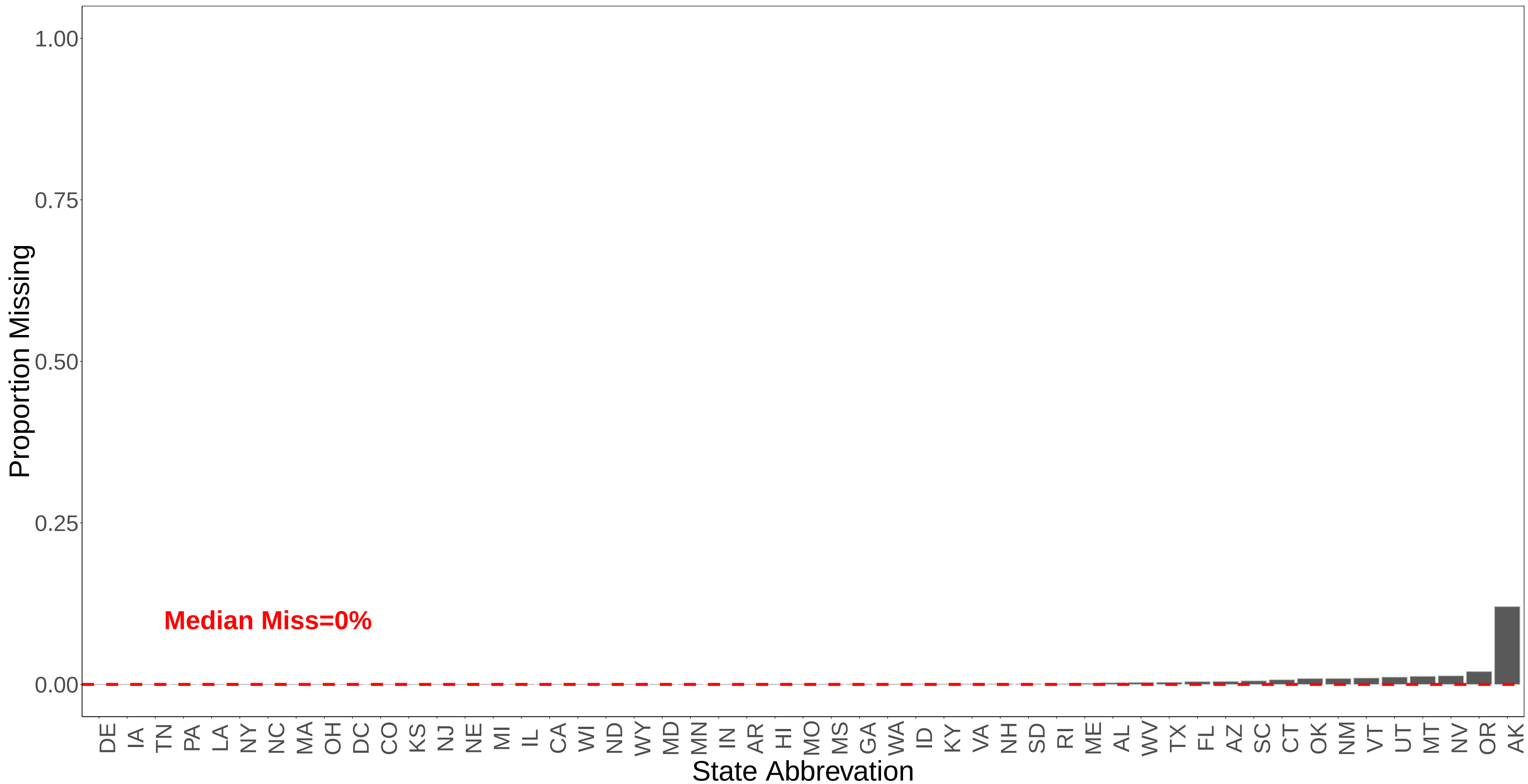
Table 1 shows that among topline estimates of proxies of data quality, the Data Trust does reasonably well. Very few members of their sample are missing a geolocated address or are missing birth year. Other measures of quality—including missing other components of birthday and registration date—are higher in terms of missingness.

Does this missingness vary across states in a way that might introduce bias intro our estimates of the effect of election laws that leverage cross-state variation? To answer this question, we first explore variation in missingness across states, then we explore whether it varies with our outcome of interest: voter turnout. Eventually, we will explore how missingness is related to identifying variation in a difference-in-difference (across states and time); however, at present that is not possible because we only have snapshot measures of missingness.

Figures 4–7 show missingness by state, following the Approach of Ansolabehere and Hersh

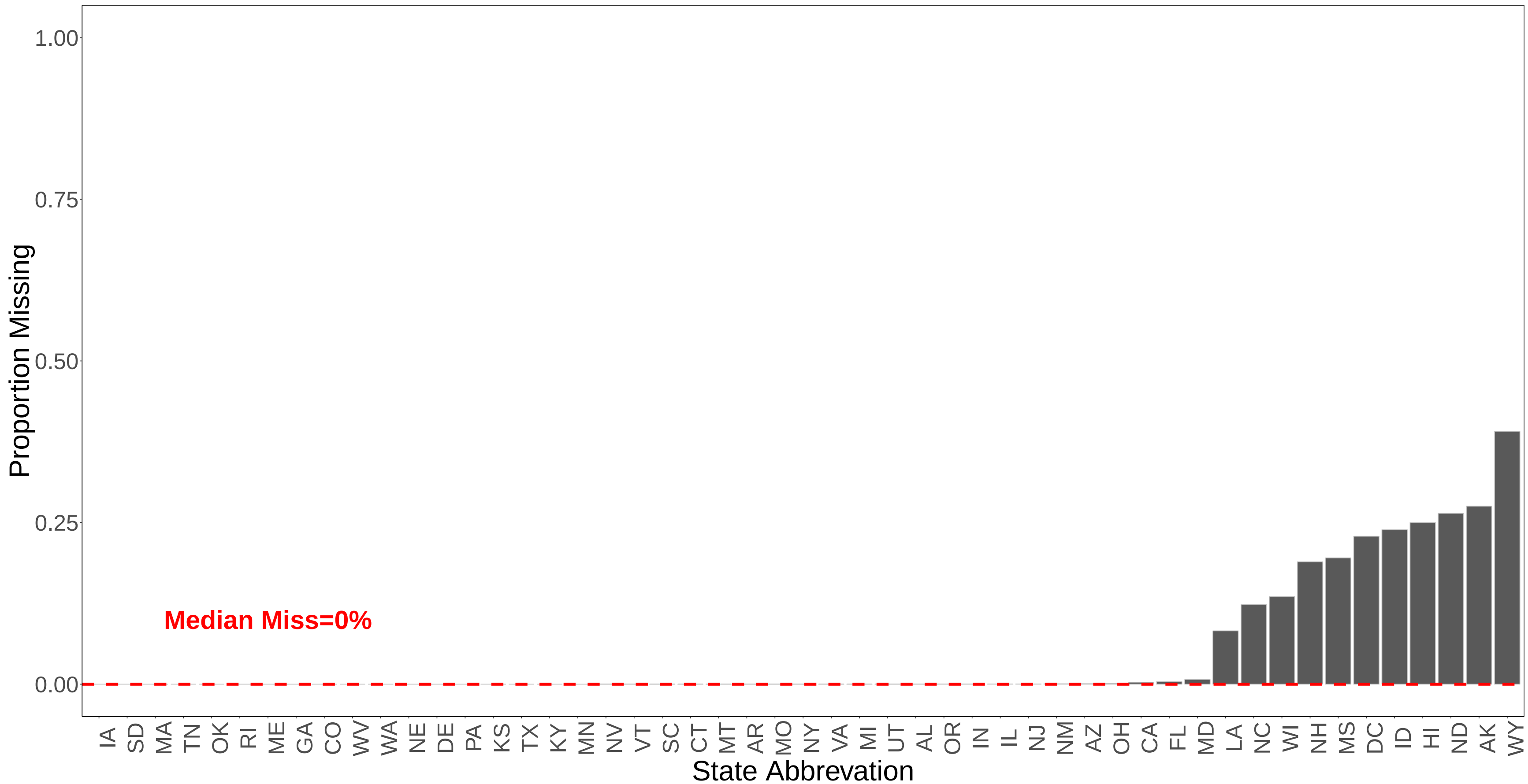
(2010) in their validation work of Catalist files. As can be seen in Figure 4, most states are not missing geocoded address. Indeed, the median state has 0% of its addresses missing. Alaska, where 12% of individuals in the file are missing geocodable addresses, bucks this pattern. However, it is more the exception than the rule.

Figure 4: Missing Geocode Address



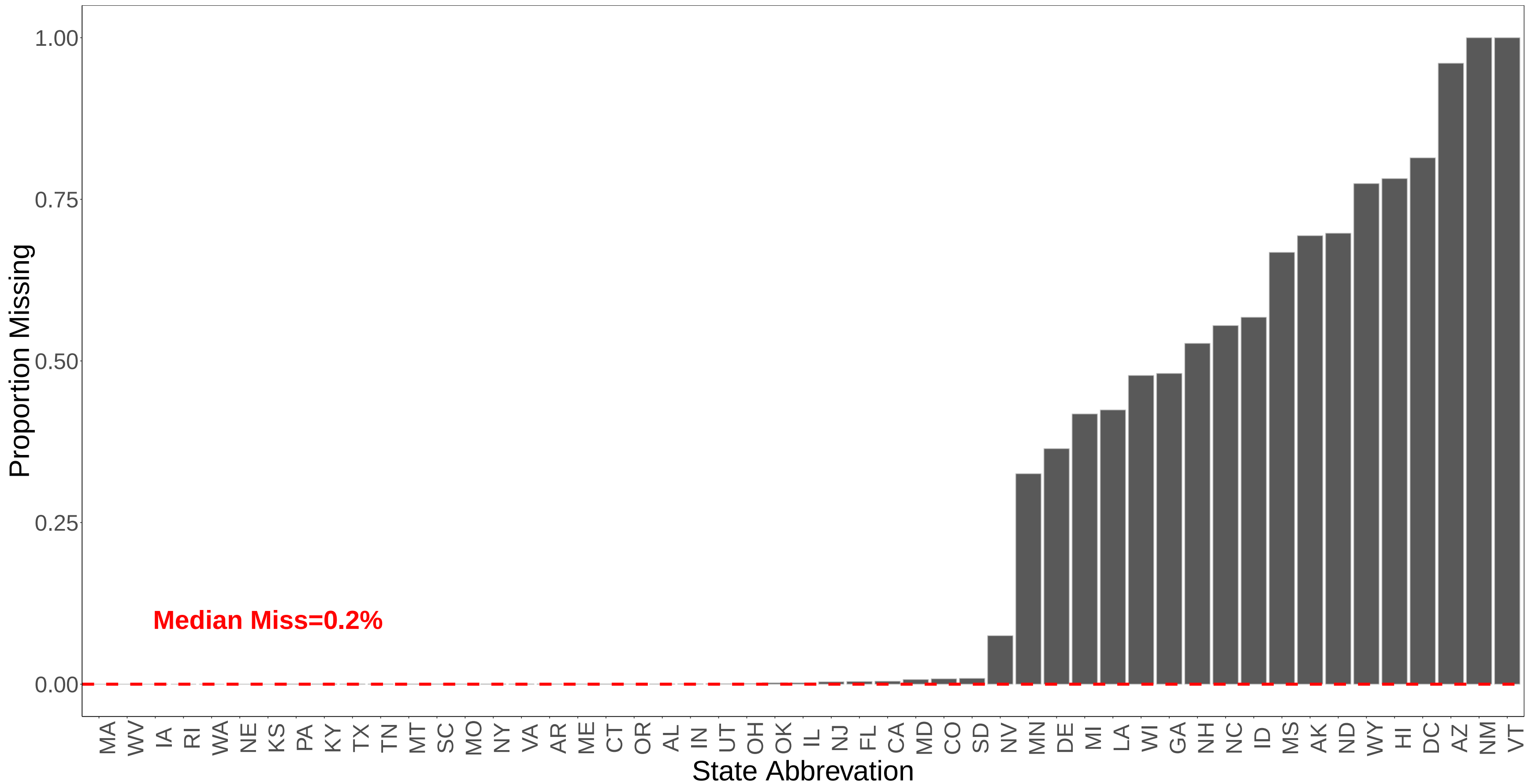
The same can be said of missing birth year. Fortunately, a dominant majority of states has, at minimum, birth year. Every state in the figure ending with Maryland has more than 99% of its observations having birth year. Starting with Louisiana, there are several states where missingness is above 5%. The highest level of missingness of birth year belongs to Wyoming: where 39.1% of individuals are missing. However, still, the level of missingness of this important variable is relatively low on the whole.

Figure 5: Missing Birth Year by State



Birth month is more frequently missing in the Data Trust files than birth year. Figure 6 shows this visually. As can be seen, a majority of states still all or mostly have birth month, as indicated by the median missing level being just above 0%. However, the number of states without this characteristic grows relative to birth year and among the states that are missing this feature, many are missing it among large chunks of the sample. Leading the way in terms of missingness are Arizona, New Mexico, and Vermont. Given laws about what can or cannot be stored in voter files in those states, virtually all individuals are missing birth month. Other states that have over 50% of the sample missing birth month include: New Hampshire, North Carolina, Idaho, Mississippi, Alaska, North Dakota, Wyoming, Hawaii, and the District of Columbia. Also of interest here is the (nearly) bimodal distribution of this variable; indeed, there is a sharp jump up in missingness from states that have all individuals with a birth month in the file and those that do not. (The jump from Nevada to Minnesota is 25 percentage points.)

Figure 6: Missing Birth Month by State



In the Data Trust files, the median state has 9.7% of its observations missing registration date. While this is higher than the other estimates provided for other data quality measures, we note that many states have a similar level of missingness on this variable. 28/49 (57%) of states (NH and ND have no registration dates because of unique electoral environments) have less than 10% missing and 38/49 (78%) have less than 20% missing. For many states, it's common for most voters to have their registration dates listed. The worst state (aside from NH and ND) in terms of reporting registration dates is Georgia (56.9% missing).[10]

[10] In the Online Appendix, we follow the lead of Ansolabehere and Hersh (2010) and look for patterns of the timing of registration (see Figure **??**). As Ansolabehere and Hersh (2010) note, there are reasons to expect that registration dates in many states may default to January 1st. We do not find strong evidence that this is occurring in many states.

Figure 7: Missing Registration Date by State

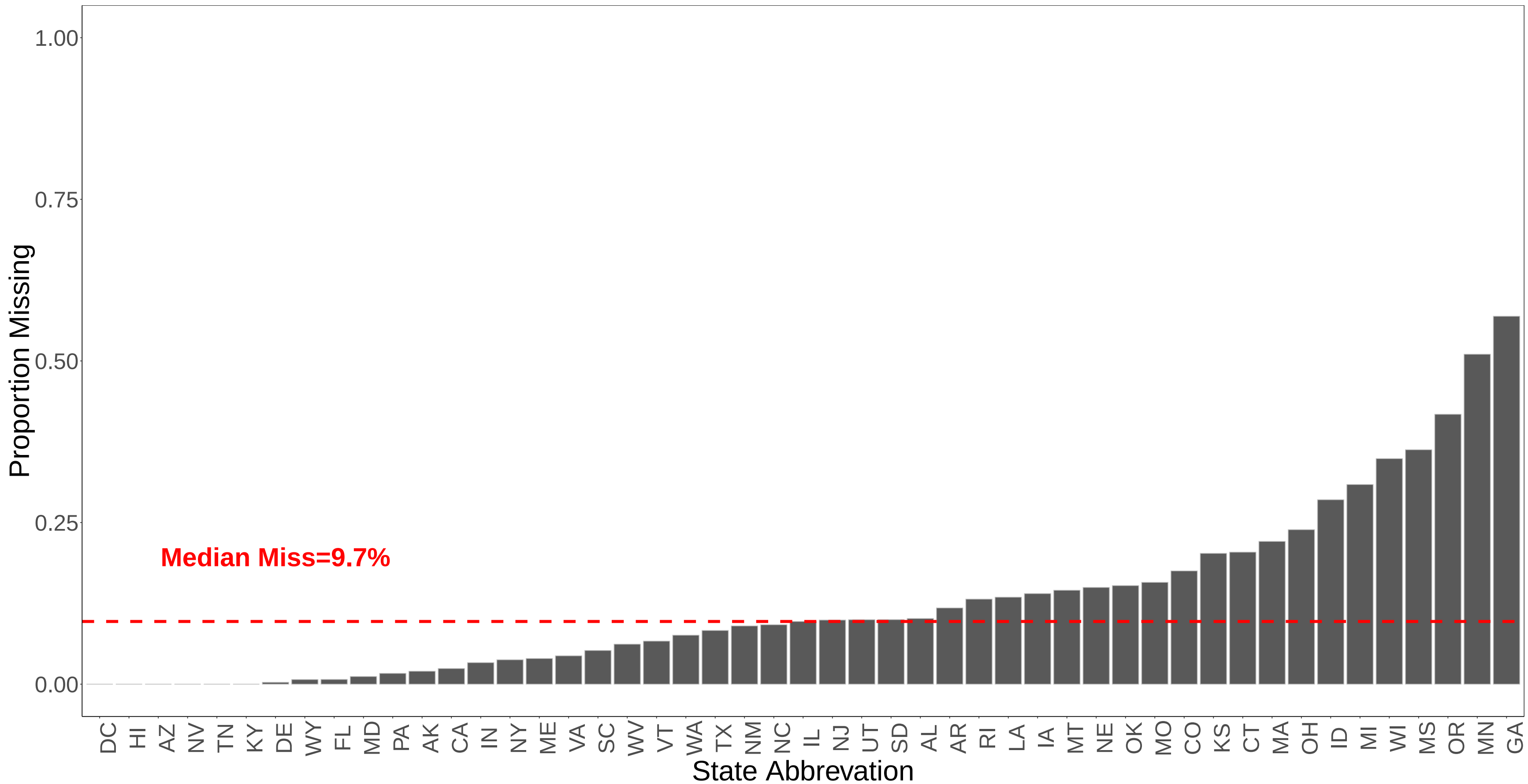
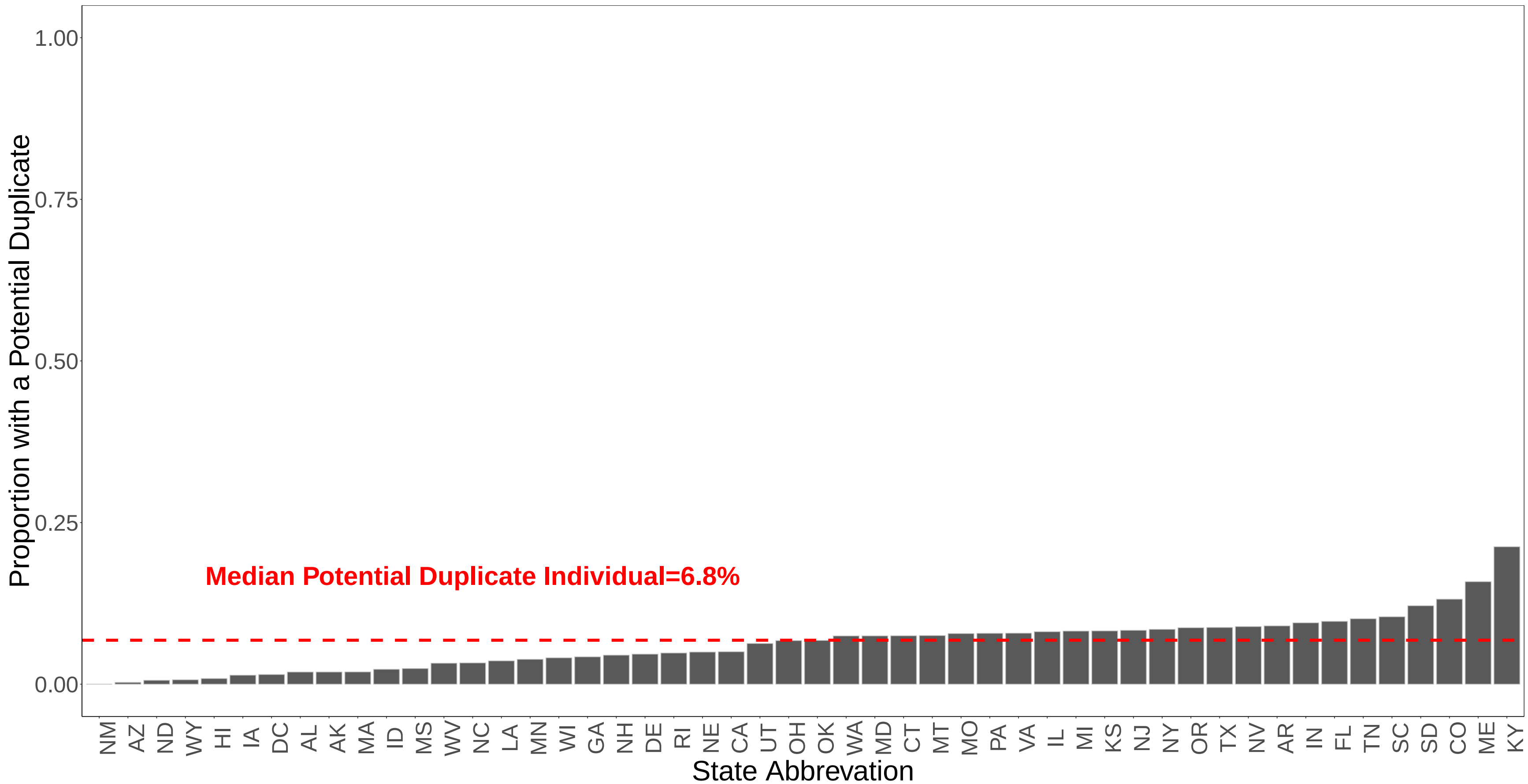


Figure 8 looks at the number of potential duplicates in the file based on a simple look for exact matches by first name, last name, and date of birth. In theory, this check could be considered to double count individuals: flagging people in their state and other states where their duplicates are. Overall, the median state has 6.8% of the file who are potential duplicates. Most states have somewhere in the same neighborhood of potential duplicates, with 44/50 states having less than 10% of registered voters fitting this category. Overall, Kentucky leads the pack, with 21.3% potential duplicates. However, this is something of an outlier.

Figure 8: Proportion of Registered Voters who Have a Potential Duplicate in the File



Ultimately, descriptive patterns like these, while informative, can only tell us party of the story of the consequences for our estimation of election law effects. Ultimately, we want to know whether data quality co-varies with election law exposure. While we do not test this directly here, we provide another check that is informative. Figure 9 plots a scatterplot between rates of voter participation (in 2014; the most recent complete election in the Data Trust data) and our several measures of missingness. As can be seen, there is virtually no relationship between a state's level of missingness of identifying information and voter turnout. The correlations for missing geocodes ($R$ =0.25), registration dates ($R$ =0.31), birth month ($R$ =0.06), birthday ($R$ =0.06), birth year ($R$ =0.15), and potential duplicates ($R$ =0.11) are all very weak. In short, these measures of data quality appear to be driven by factors unrelated to turnout. While not definitive proof of the quality of the Data Trust data, this is assuring.

Figure 9: Missingness of Key Characteristics and Voter Turnout

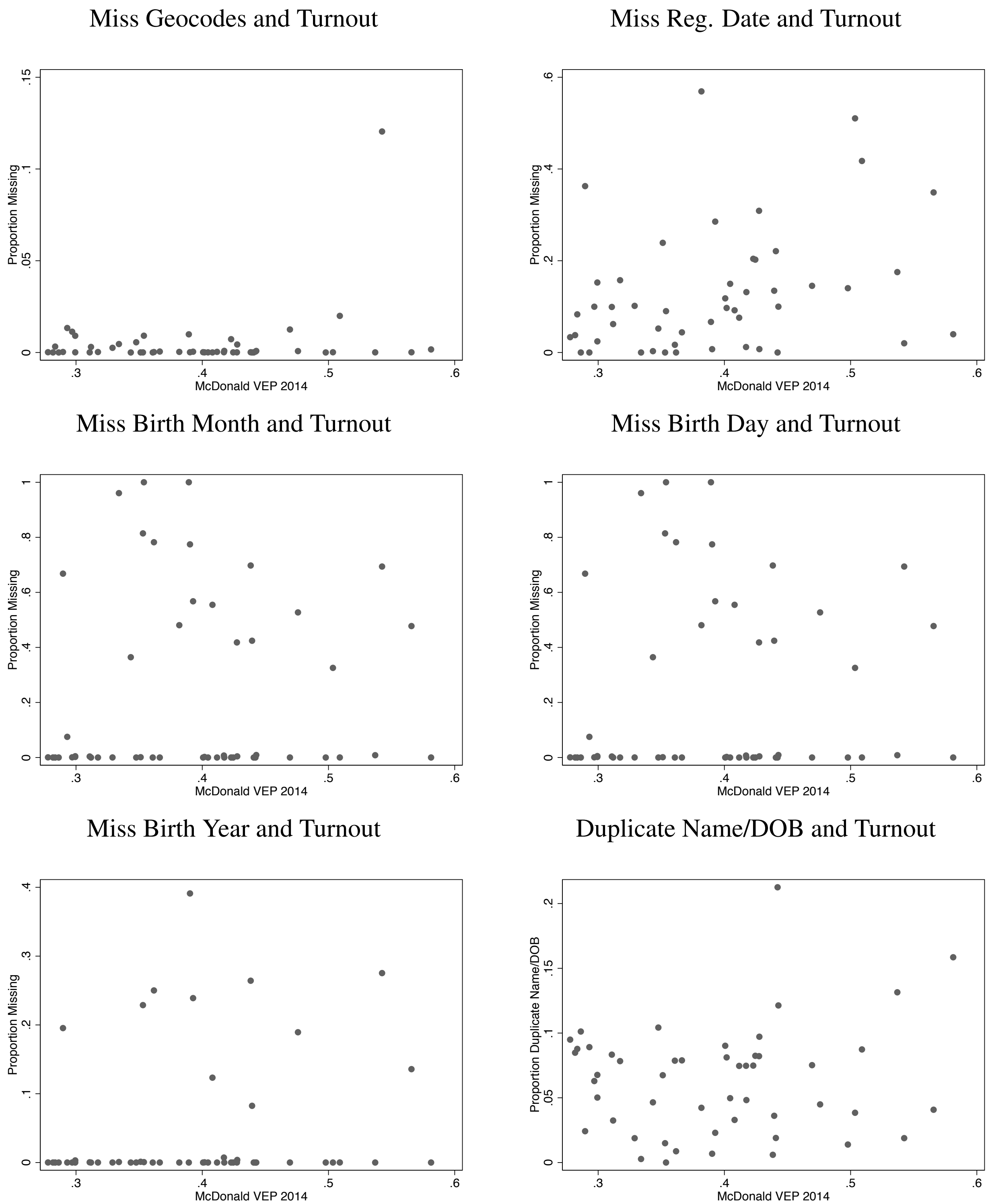



Figure 9 compares missing rates for several key variables with voter turnout (VEP, 2014). $\mathrm{R}_{geo} = 0.25, R_{regdate} = 0.31, R_{bmonth} = 0.06, R_{bday} = 0.06, R_{year} = 0.15, R_{duplicate} = 0.11$

## Discrepancies in Voter Turnout Numbers

One of the biggest questions about the quality of voter registration lists involves their ability to approximate voter turnout numbers historically. As voter files age, they, they face a number of issues related to voter purges. Does this decay restrict our ability to construct panels from contemporary snapshots? We consider this question here.

Figures 10–16 plot the discrepancy in vote counts (scaled by the voting eligible population) between the Data Trust's estimates of historical vote counts and those provide by McDonald. As a benchmark, we use McDonald's highest office estimate as it is available for all states across time, whereas the total ballots counted has missingness over time and, as McDonald notes, the total ballots measure often actually is the highest office measure.

As a general rule, nationwide snapshots perform better in more recent years. The 2017 snapshot misses turnout in the median state by 1.0 percentage points (equivalent to just over 17,000 votes; see Figure OA3 in the Online Appendix). Over time, these discrepancies steadily grow, until in 2004 and the misses are quite large—as larger as 18.5 percentage points in the median state in 2004.

While this looks bad at first, we think it important to note that the usefulness of historical turnout from contemporary files depends on what one wants to do. If the goal is to accurately measure historical turnout levels, contemporary voter files clearly present a problem. However, the key question in the study of election law effects is the extent to which misses co-vary with treatments of interest. That is, if states were to systematically differ in how much they miss voter turnout numbers in a way that was related to the adoption of election laws, it may introduce bias into our estimates.

On thing that is clear across Figures 10–16 is the fact that misses in many states are quite similar within a given year. This can be seen by looking at the level of variation in misses across states. Though the eye is naturally drawn to the levels and tails of these graphs, we think it important to note that in many states turnout misses are quite similar. Turnout misses have the appearance of a relatively consistent unit shift across states. While historical estimates in contemporary files miss turnout by quite a lot, it does so fairly uniformly across states. This can be seen in Figure OA1 in the Online Appendix, which plots the distribution of misses in turnout over time. While the gap between official turnout estimates and the Data Trust estimates, it does so fairly uniformly: with most states decaying at a similar rate.

Another way to see this similarity in turnout decay over time is to look at the relationship between turnout counts and official counts. As can be seen in Figure 17, there is clearly drop-off over time in voter turnout relative to McDonald's estimates; however, this drop-off is fairly uniform. While turnout back in time does not get to actual levels—with the exception of 2014—the relationship between these two variables is still very strong: with the Pearson's R being above 0.99 in all years. In short, it doesn't appear that there is differential levels of decay across the states.

Figure 10: Difference between McDonald and Data Trust Vote Counts Scaled by VEP 2014

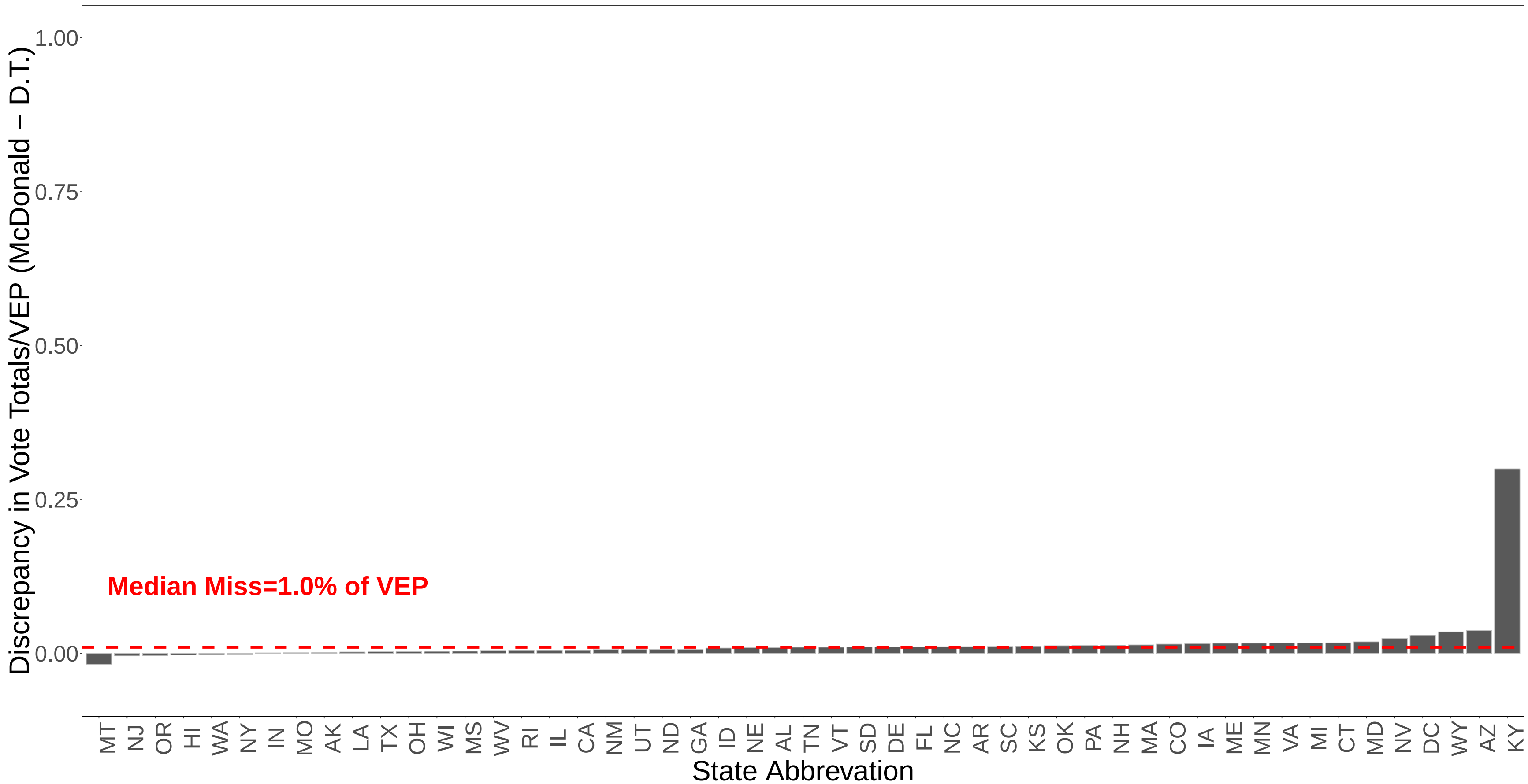


Figure 11: Difference between McDonald and Data Trust Vote Counts Scaled by VEP 2012

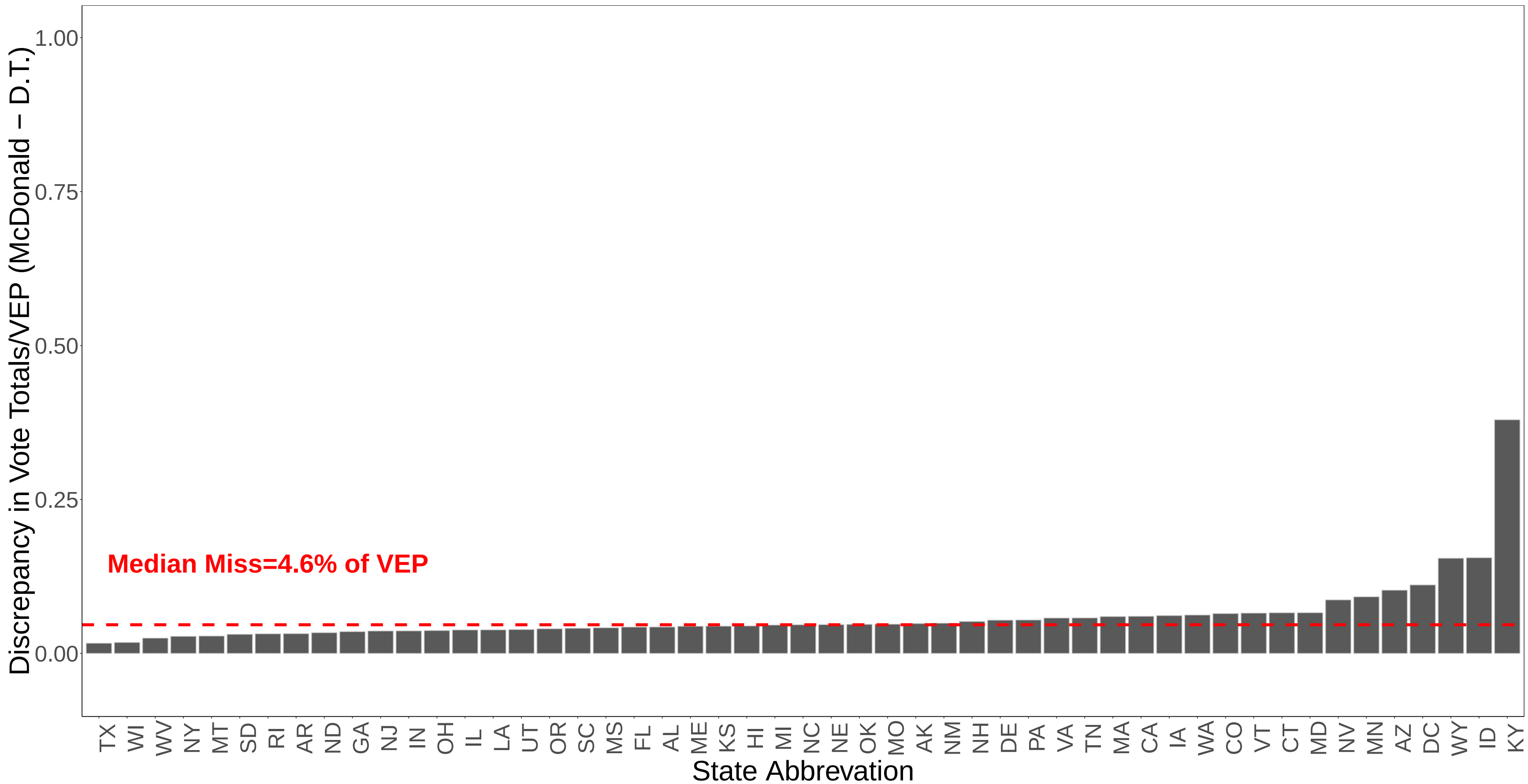


Figure 12: Difference between McDonald and Data Trust Vote Counts Scaled by VEP 2010

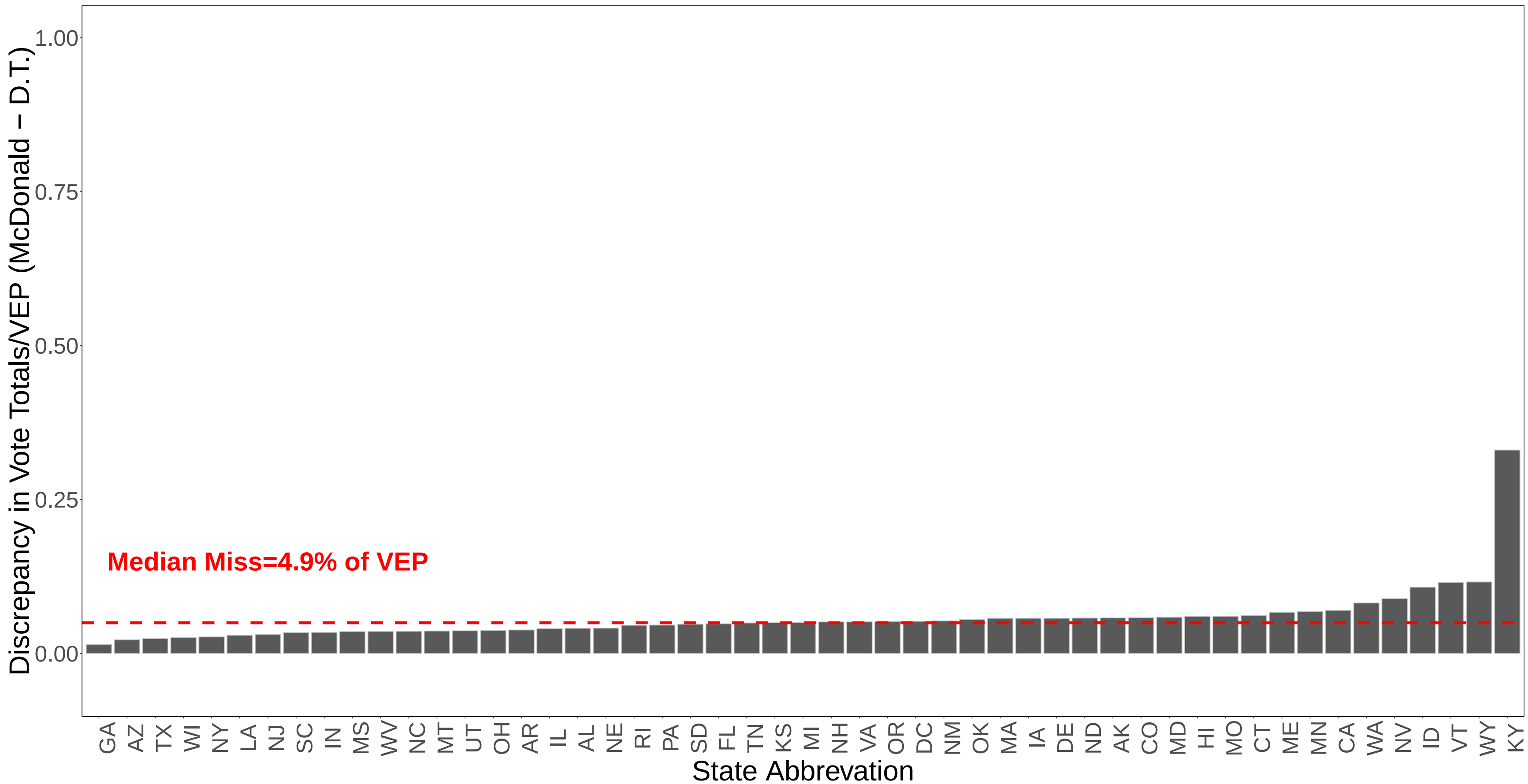


Figure 13: Difference between McDonald and Data Trust Vote Counts Scaled by VEP 2008

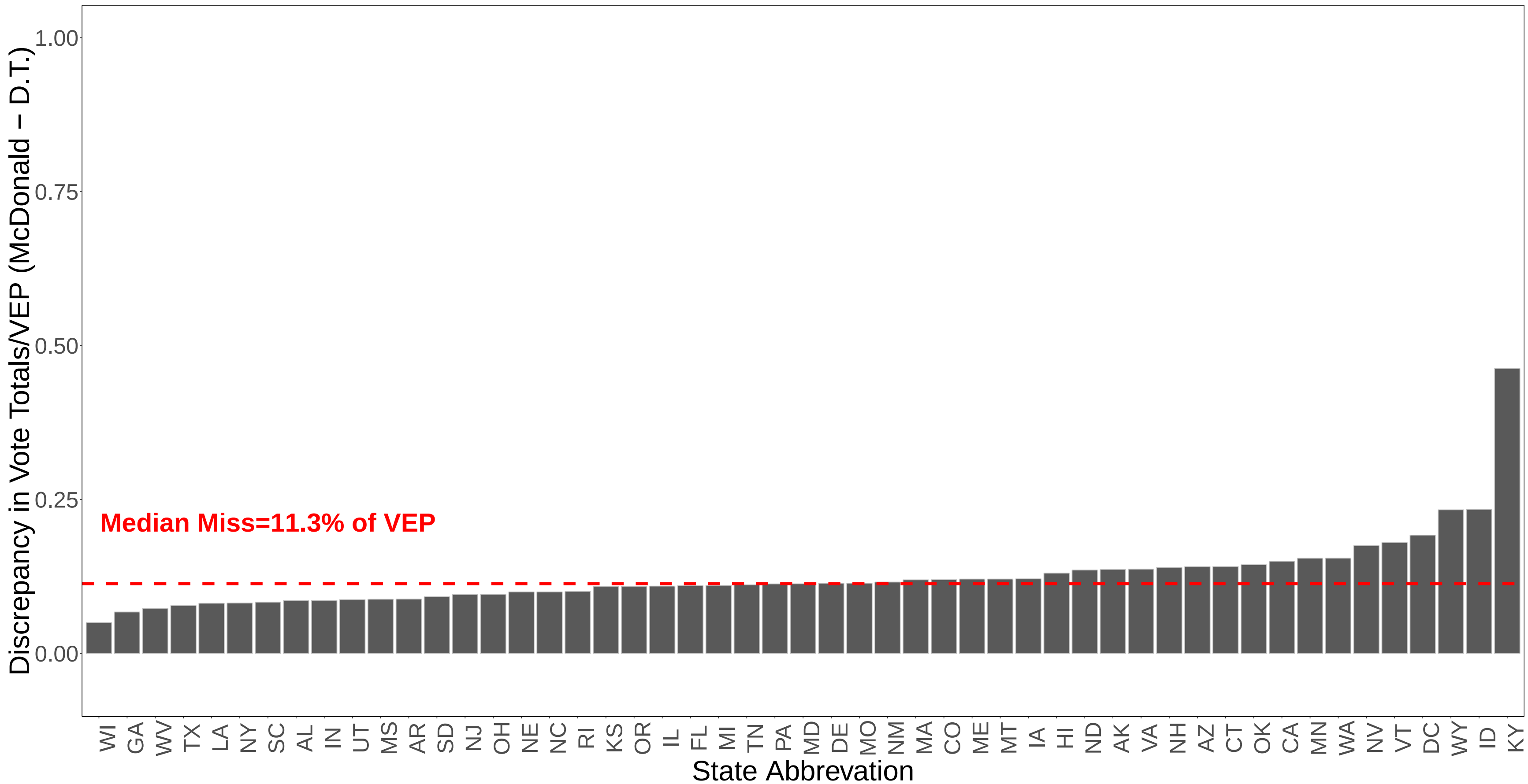


Figure 14: Difference between McDonald and Data Trust Vote Counts Scaled by VEP 2006

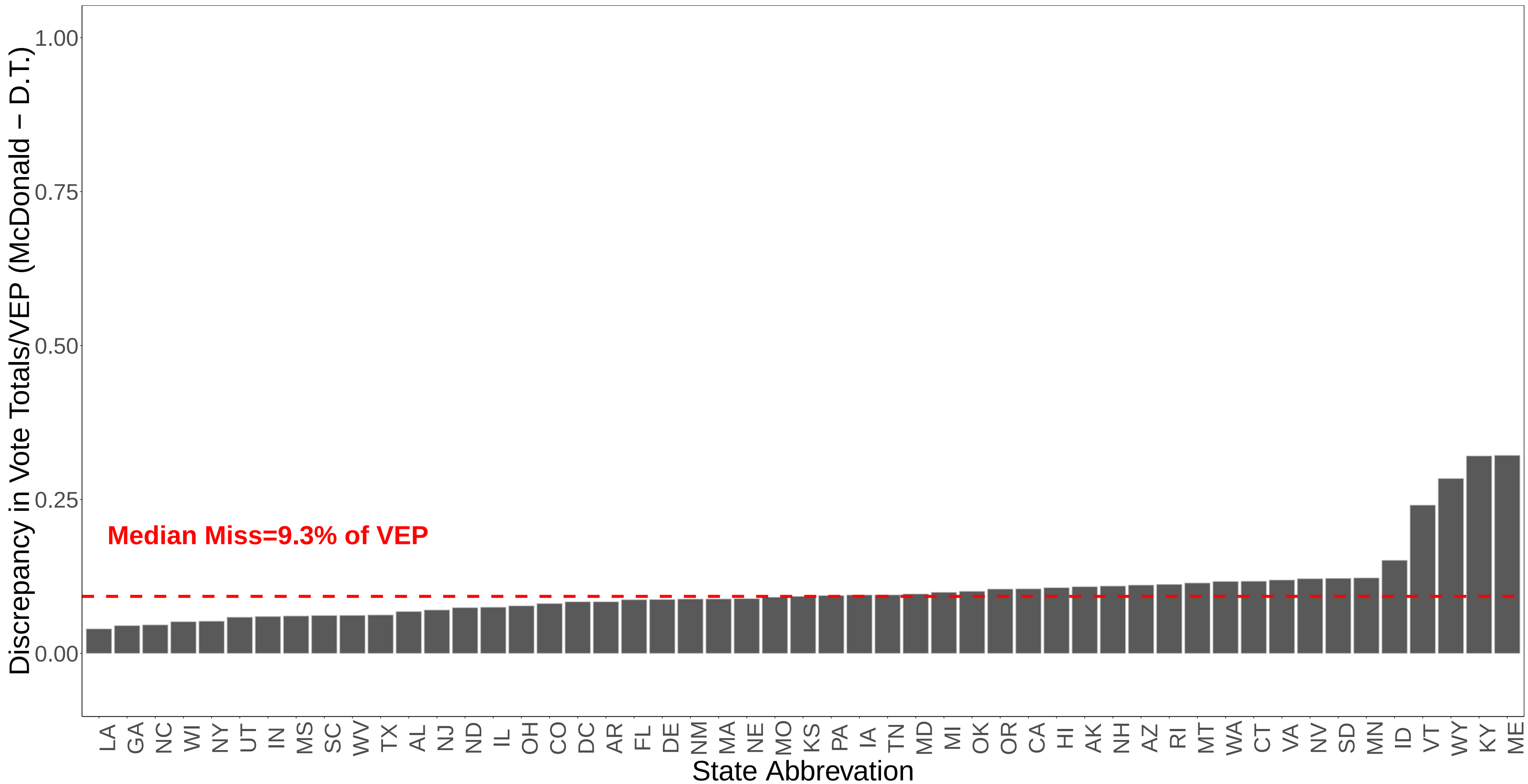


Figure 15: Difference between McDonald and Data Trust Vote Counts Scaled by VEP 2004

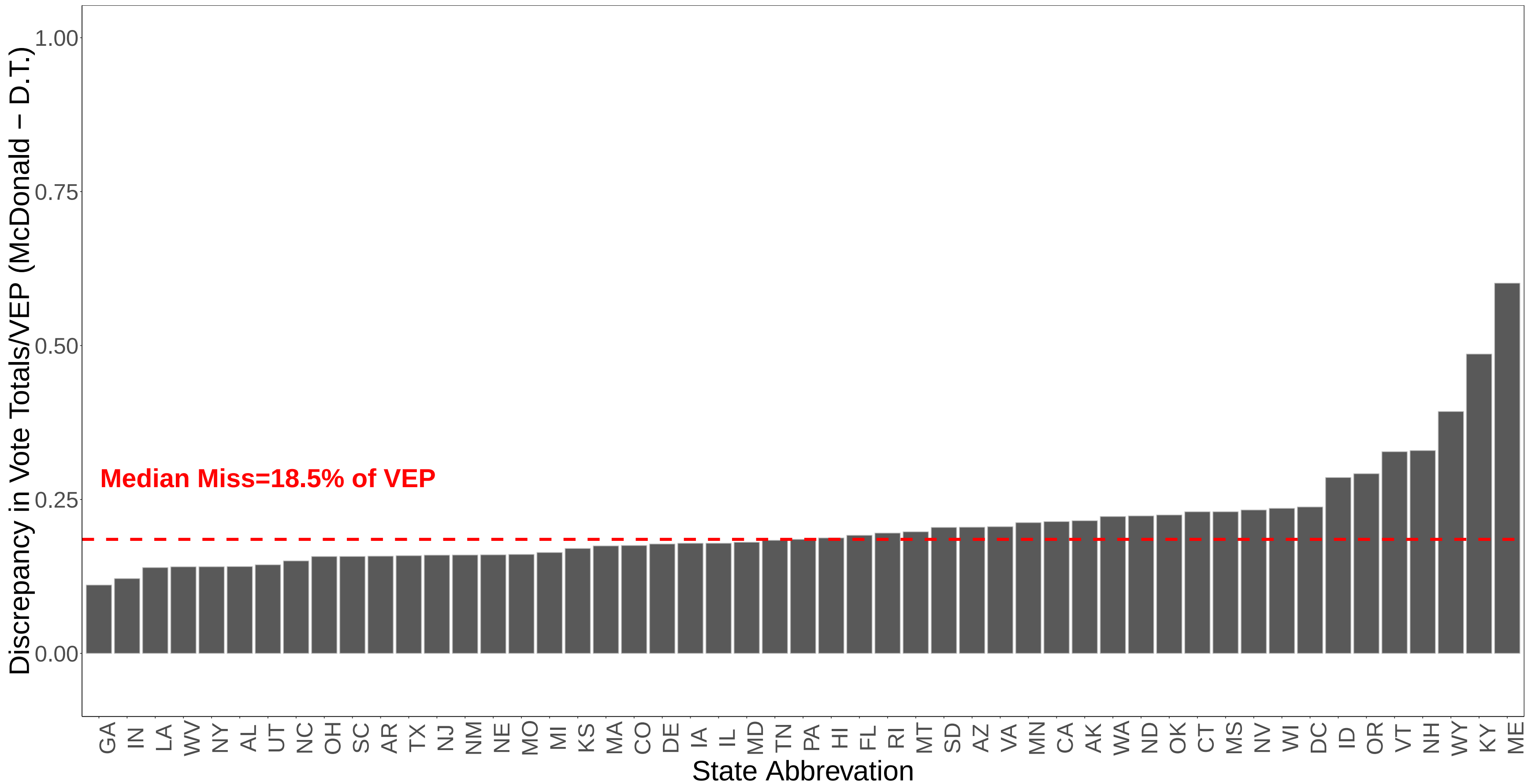


Figure 16: Difference between McDonald and Data Trust Vote Counts Scaled by VEP 2002

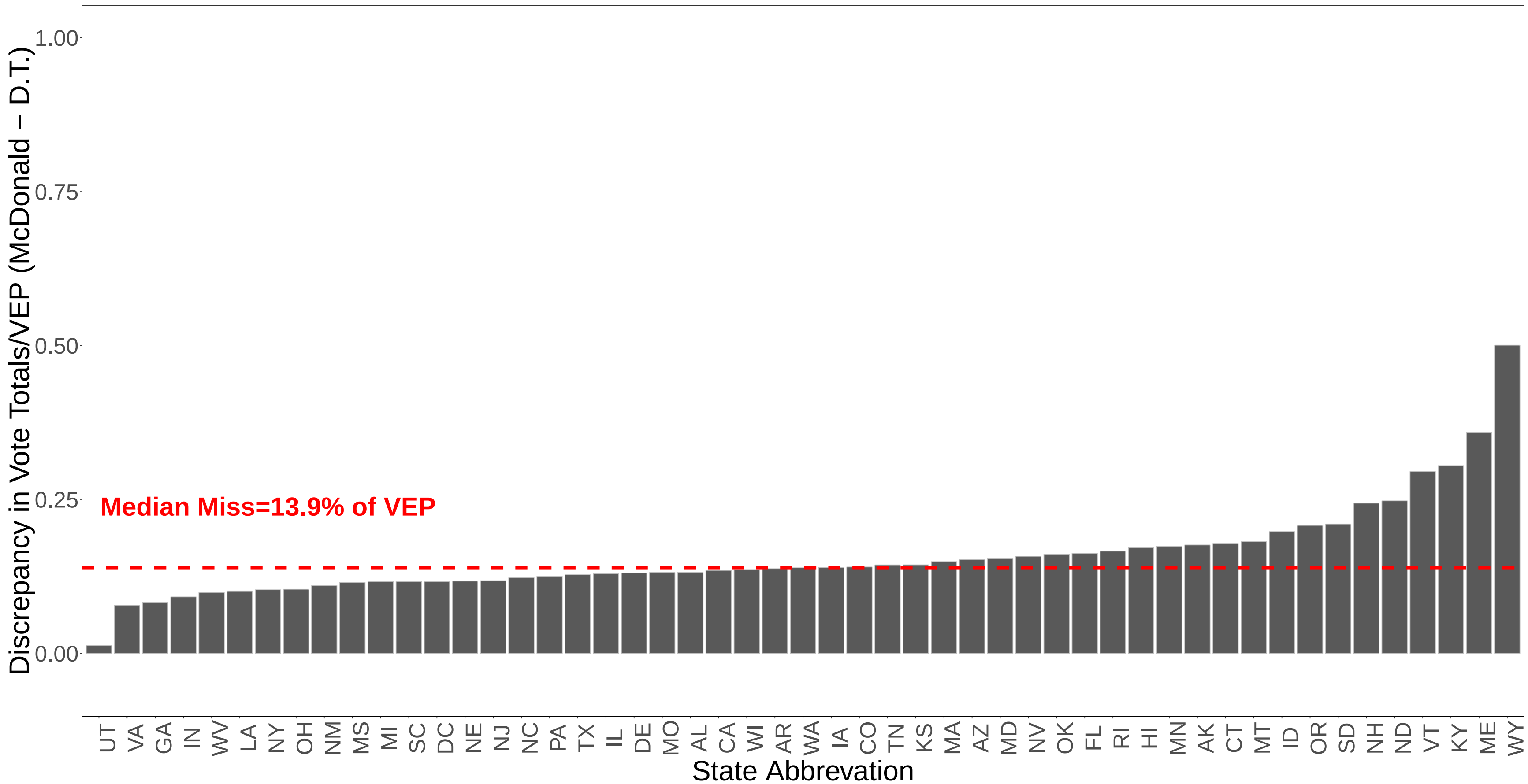


Figure 17: Vote Counts in the Data Trust Files Benchmarked to McDonald Counts

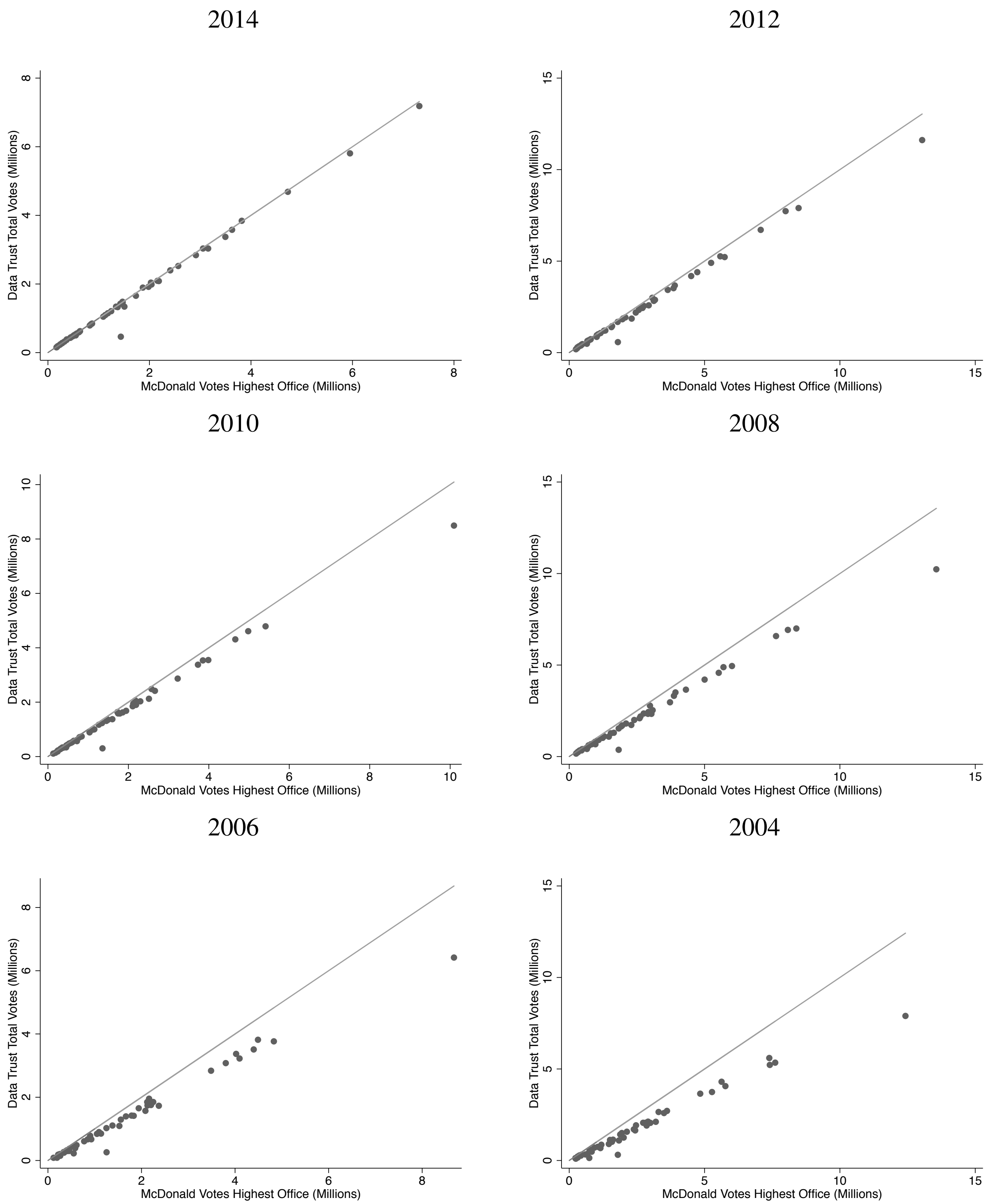



Figure 17 plots the relationship between official highest office vote count numbers via the U.S. Elections Project (on the X axis) and the Data Trust turnout counts (on the y axis) at the state level, broken by year. The line represents where points would be if turnout in both sources were exactly the same. Yearly correlation coefficients: 0.996 (2014), 0.997 (2012), 0.995 (2010), 0.994 (2008), 0.993 (2006).

One final way to look at whether turnout decay affects electoral law estimates is to formally estimate the effect of said laws on the amount of discrepancy from official estimates. This involves substituting turnout with the turnout discrepancy (at the state-year level) as the dependent variable. This allows us to look at a very similar model (and various modifications to that model) that many run: a difference-in-difference.

To do so, we use voter identification laws as our case study given their salience in the literature. We run several specifications of a difference-in-difference model of this type. The effects are broken by whether the treatment of interest is a non-strict photo identification law (on the left) or a strict photo identification law (on the right). Across both panels, *Model 1* is a classic two-way fixed effects model frequency weighted by the number of registered individuals in the state year. *Model 2* is the same specification as Model 1, but unweighted. *Model 3* adds a linear time trend interacted with the state fixed effects to Model 1. *Model 4* is the same as Model 3, minus frequency weights. All models follow previous practice and cluster standard errors at the state level.

Figure 18 displays the results from said models. As can be seen, voter ID laws seem to have very little effect on voter turnout decay. The average difference between official estimates and those from the Data Trust files (scaled by VEP in the state-year) does not vary along voter ID. The average effect size across model specifications is small (0.8 percentage points) and the average p-value is large (p=0.28). One of the models (model 2) for strict voter-ID—clears the statistical significance threshold (-0.029). However, the estimates for this model remains quite small and is not robust to other variations in model specification that we run. That is, in a standard model that one might use to estimate the effects of voter ID laws, there is balance in the difference between official and contemporary voter file estimates. This would provide further evidence that decay might not be influencing election law estimates as some have previously speculated. We recommend that scholars run a test like this before using historical turnout from contemporary files in estimating the effect of election laws. If one is worried about potential differential decay across subgroups, models with decay among those subgroups could also be run.

Figure 18: Difference-in-Difference Estimates of the Effect of Voter ID Laws on the Discrepancy Between Official and Data Trust Turnout Estimates (2006-2014)

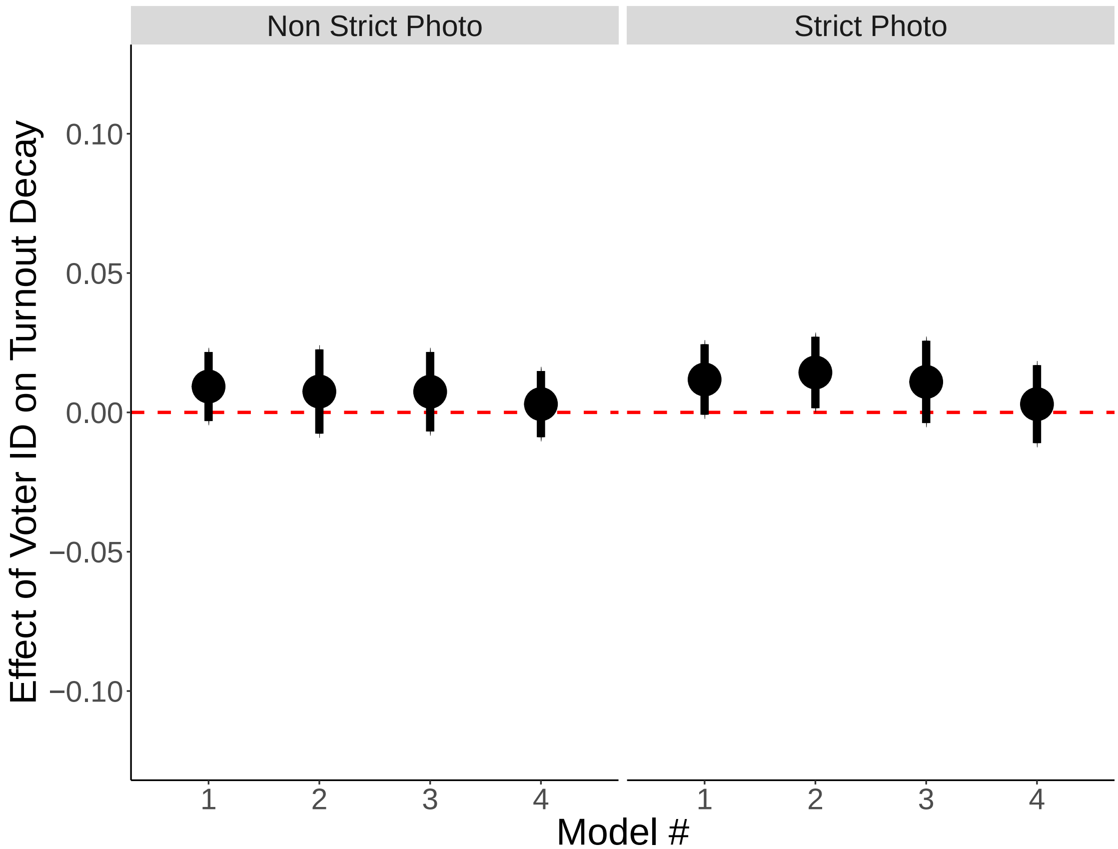



Figure 18 shows the state-year level difference-in-difference estimates of the presence of voter ID laws (strict and non-strict) on the discrepancy between the Data Trust's turnout numbers from the 2017 snapshot and official turnout counts from the U.S. Elections Project from 2006-2014. *Model 1* is a classic two-way fixed effects model frequency weighted by the number of registered individuals in the state year. *Model 2* is the same specification as Model 1, but unweighted. *Model 3* adds a linear time trend interacted with the state fixed effects to Model 1. *Model 4* is the same as Model 3, minus frequency weights. All models follow previous practice and cluster standard errors at the state level. N from left to right: 601634974, 255, 601634974, 255, 601634974, 255, 601634974, 255. The differences in sample size are entirely attributable to whether a model has frequency weights or not. All models cluster standard errors at the state level.

While turnout decay is certainly a potential problem, we view it as one that is not insurmountable (in many election law cases). If scholars are worried that purges from the file, we recommend that they measure the discrepancy between the counts in their file and those of McDonald and include this as a time-varying state-level covariate.

## Methods for using a national voter file to study election laws

Having discussed some of the potential data issues facing analysts who want to use nationwide voter files to study the effect of election laws, we now turn our attention to outlining methodological approaches to this topical area. The structure of nationwide voter files allows for more flexibility in designing studies than many other commonly used data sources in the study of election laws and rules. In this section, we lay out some of the unique designs that are facilitated by the use of national voter files.

With a nationwide voter file, we recommend tests of at least five types that can be used in many different circumstances to study interventions that are assigned at levels of geographic aggregation. These designs leverage the individual-level nature of the data, the over-time structure of the data, and the size of the national voter file to expand the universe of possible research designs. Our recommendation is that analysts using national voter files be as thorough as possible, given the difficulty of identifying the causal effects of election laws. Each test has its own assumptions, so we outline these here. As always, we recommend that analysts thoroughly plan their designs and register pre-analysis plans with appropriate repositories before conduct any analyses.

In all cases, the data availability and quality issues outlined above should guide the choice of analytic approaches. Despite the limitations of voter file data, though, individual-level national voter files provide important analytic power through both their size and the over-time nature of the data. As such, all of the proposed analytic methods below can be used to conduct comparisons on fine-grained groups such as narrow age ranges, intersections of age and race, and intersections of age, race and partisanship. Survey-based approaches would lack the power to detect effects in many of these small subgroups, even though they are often the units of analysis most interesting to analysts and the ones central to legal battles over election laws (Grimmer et al. 2018).

### Difference-in-difference

As a first design, analysts can use a canonical two-way fixed effects/difference-in-difference estimator to estimate the effect of residing in a jurisdiction that implements a new election law. The treatment indicator in these models is whether a respondent's state has implemented the law or policy being studied at the time of the election. As an individual-level analysis, the dependent variable is whether a voter is coded as voting in that year's election. This model assumes that states that adopt into new election laws have parallel trends in voting to those that do not adopt the laws.

While these difference-in-difference models have become common in the study of election laws using survey data or other administrative data, the national voter files provides the opportunity to fit models to narrower subgroups of the population than many other data sources are able to precisely estimate effects for. Of course, analysts should take care to pre-register subgroup hypotheses and to consider tests for multiple comparisons before conducting such analyses

## Difference-in-difference with flexible time trends

To supplement these difference-in-differences models, in our second set of models we add flexible time trends for each state. Like the two-way fixed effects models, these absorb all observed and unobserved factors that remain constant within states and are shared within certain years, while also accounting for yearly trends that vary across states (e.g. the natural trajectory of turnout across states). The inclusion of state-specific time trends allows us to relax the sometimes tenuous parallel trends assumption key to difference-in-differences specifications. Here our identifying assumption is that our outcomes in states separate by treatment deviate from common year effects by following the linear trend captured by the interaction term. Under this assumption, identification comes from sharp deviations from otherwise smooth state-specific trends. The assumption behind this approach is inherently untestable, although it is considered to be stronger than one required in a model with just state and year fixed effects (Angrist and Pischke 2014).

## Difference-in-difference with individual fixed effects

To leverage the individual-level panel nature of voter file data, we recommend estimating models that include individual fixed effects. By leveraging individual changes in turnout over time, this model will allow us to rule out a host of unobservable characteristics that are potentially not captured by the two-way fixed effects or state time trends. However, including individual fixed effects comes at the cost of the exclusion of controls that analysts might want to include in other difference-in-difference models. For example, the Data Trust data contains many modeled and unmodeled variables for each voter, but most of these variables do not vary over time, and thus they cannot be used as controls in the same model as the individual fixed effects. This modeling approach assumes that changes to voter ID laws are orthogonal to changes at the individual-level.

## Matching

Large, individual-level datasets are well-suited to a variety of matching analyses. In the study of voter ID laws, this approach was specifically suggested by Highton (2017, 159, 164) as a supplement to the classic difference-in-difference specification.

One limitation to using voter files for matching is that the quality and/or number of right-hand-side variables to use in finding matches. While commercial voter files can include many hundreds of variables merged from commercial sources or modeled from other data, these variables can have considerable missingness and may not be appropriate for valid matching analyses. Even basic demographic variables such as race/ethnicity and income must be modeled and are thus potentially prone to more error than in survey data.

However, the size of the national voter file provides an opportunity to conduct very precise exact matching on the variables that are measured with precision in voter files. As an initial step in matching analyses, we recommend matching on individuals' past turnout behavior, going back to the oldest year in which the data are reliable (for our Data Trust voter file, 2006), and age, which the voter files measure with good precision. These matching analyses will provide close matches without relying on modeled data.

In additional matching tests, we suggest building in both registered and modeled partisanship, conducting the analysis separately for party registration and non-party registration states if possible to check whether including modeled partisanship estimates are biasing the results. Other modeled and unmodeled covariates from voter files could be added to matching models to further increase the precision of the matches, after appropriate checks for data quality. These models assume that any time-varying heterogeneity not soaked up by the state and year fixed effects will be captured by these observable controls.

## Synthetic controls

Individual-level voter file data also allow for the use of the synthetic control method, which allows for time-varying individual-level heterogeneity (Abadie, Diamond and Hainmueller 2015). While the most common use of the synthetic control method estimates state-level average treatment effects, the method can be expanded in the case of the large data in the national voter file to a variety of applications. For example, a subgroup analysis could be constructed for individual racial groups that constructs synthetic control states by weighting over control states to match pre-implementation trends for each racial group separately. These synthetic control tests provide an additional method to assess the validity of cross-state comparisons.

## Regression discontinuity designs

Finally, the national voter file provides the opportunity for analysts to conduct geographic regression discontinuity designs. Because voters in voter files are identified by their home addresses, they can easily be geocoded (and commercial voter files such as our Data Trust file have already geocoded voters). These voters can be used to estimate designs that use geographic borders around which election policies are different (whether state, county, etc.) to estimate treatment effects of exposure to laws being studied. Geographic regression discontinuity designs require important assumptions to produce valid causal estimates of treatment effects, so analysts need to take care to defend their assumptions in specifying these designs (Calonico, Cattaneo and Titiunik 2014; Keele and Titiunik 2015; Keele, Titiunik and Zubizarreta 2015). Nevertheless, the scale and geolocated nature of national voter files provides a greater opportunity for valid and precise geographic RDDs than other extant data sources.

## Differential registration bias

One of the challenges that is inherent when one is working with voter files is that fact that they implicitly condition on registration. That is, in pulling from the universe of registered voters in the state, one is getting a skewed sample, one that does not include non-registered voters. This can induce bias if the treatment of interest affects registration rates substantially (Nyhan, Skovron and Titiunik 2017).

Correcting for differential registration bias is difficult. At the individual level it is hard to know what dataset one should use to supplement the voter file. That being said, this problem is not insurmountable. In line with Nyhan, Skovron and Titiunik (2017), we note that it's important to check the extent of differential registration bias in the sample of study.

**Checking for Differential Registration Bias**

To assess the extent of differential registration bias (Nyhan, Skovron and Titiunik 2017), we recommend that researchers using voter files to estimate election law effects look explicitly at registration rates as an outcome. This will allow them to see whether conditioning on registration in an individual-level voter file has any effect on voter turnout estimates.

Fortunately, doing so with a nationwide file is relatively straightforward. Doing so necessitates that one collapse the dependent variable (registration) down to a higher level unit as there are very few non-registrants in most consumer files and these individuals are unlikely to be representative of non-registered citizens as a whole. Hence, running individual-level regressions with registration as an outcome in nationwide voter files is unpalatable. One could collapse registration down to the state-year and look for effects like one does with turnout. Or, alternatively, as registration occurs more regularly than voting, one could look more fine grained at registration rates by calendar day. This would be conducive to a temporal regression discontinuity approach, where one looks around the timing of when election laws are implemented. Regardless of the unit one chooses, once the collapsing is done, registration numbers can be scaled by a reasonable denominator, like the voting eligible population.[11] One can then run an analogous model to what they are running for voter turnout on registration instead. This process could be repeated for subgroups of interest, with the only shift being a change in the denominator.[12]

Figure 19 shows this approach for a special case: voter identification laws—a hotly discussed and debated election law. As above, it shows estimates across several model variants that researchers might, and have, use(d) to estimate the effect of election laws using cross-state variation. The effects are broken by whether the treatment of interest is a non-strict photo identification law (on the left) or a strict photo identification law (on the right). Across both panels, *Model 1* is a classic two-way fixed effects model frequency weighted by the number of registered individuals in the state year. *Model 2* is the same specification as Model 1, but unweighted. *Model 3* adds a linear time trend interacted with the state fixed effects to Model 1. *Model 4* is the same as Model 3, minus frequency weights. All models follow previous practice and cluster standard errors at the state level.

As can be seen, in the case of voter identification laws, the evidence for an effect on registration is quite weak. Across all 8 model specifications, the estimates are all small—averaging -0.4 percentage points—and not close to being statistically significant (the average p-value across model specifications is 0.6). This lack of significance is not for a lack of statistical power. Indeed, these null effects are quite precise—allowing us to confidently rule out very small to modest effects, particularly in the models with individual weights (1 and 3). But, even in the unweighted models (2 and 4), the effects are quite precise.

[11] An analysis where the day is the unit of analysis probably necessitates that one look at the number of people registering, as the denominator in this setup is less clear.

[12] Since VEPs by subgroups are hard to come by, one could use subgroup population estimates.

Figure 19: Difference-in-Difference Estimates of the Effect of Voter ID Laws on Registration Rates (2006-2014)

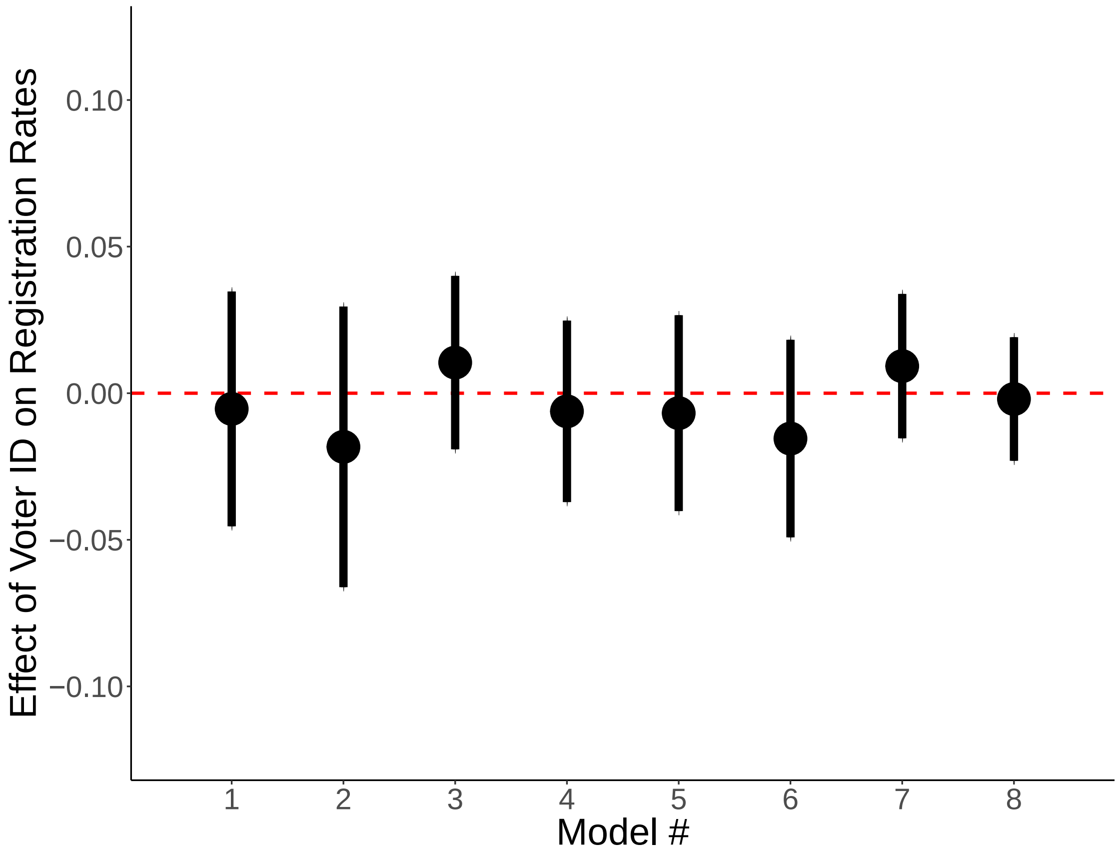



Figure 19 shows the difference-in-difference estimates of the presence of voter ID laws (strict and non-strict) on registration rates between 2006-2014. *Model 1* is a classic two-way fixed effects model frequency weighted by the number of registered individuals in the state year. *Model 2* is the same specification as Model 1, but unweighted. *Model 3* adds a linear time trend interacted with the state fixed effects to Model 1. *Model 4* is the same as Model 3, minus frequency weights. All models follow previous practice and cluster standard errors at the state level. N from left to right: 601634974, 245, 601634974, 245, 601634974, 245, 601634974, 245. The differences in sample size are entirely attributable to whether a model has frequency weights or not. All models cluster standard errors at the state level.

These results suggest that in the case of some electoral laws—like voter ID—the risk for differential registration bias may not come into play. To be clear, this may not be true for all electoral laws. Some may be more likely to have effects on registration than others, thus necessitating adjustments/sensitivity tests suggested in the literature (Nyhan, Skovron and Titiunik 2017). Our recommendation is that researchers explicitly look for differential registration bias when using nationwide voter files to estimate the effects of election laws. If one is worried about potential differences in registration rates across subgroups, they could easily calculate state-year registration rates for those subgroups and re-run these models again. These checks are not perfect, but they shine light on the extent to which common research designs may be influenced by differential registration patterns.

# Conclusion

In this paper, we explored the potential issues with using nationwide voter file data to estimate the effect of election laws. Despite having potential issues, we have argued that many of these may actually not be so large. Contemporary nationwide snapshots do well in matching to known quantities nationally and across states and some of the other potential data issues may not bias election law estimates, like the effect of voter ID laws.

We also discussed potential new methodological techniques that can be paired with these fine-grained datasets. We outlined the strengths, weaknesses, and assumptions of various modeling techniques: individual matching, individual fixed effects, and synthetic controls, to name a few. These methods have been underutilized in the election law literature. They offer to shine insights into debates that have, for too long, relied on a the canonical two-way fixed effects specification that has dominated the literature.

In short, nationwide voter files offer tremendous value in addressing some of the problems that survey data cannot. These datasets are not perfect, and researchers should exercise care before charging headlong into analyses estimating the effect of election laws. However, using these datasets to achieve this goal seems reasonable when paired with the checks we have recommended here. The next logical step, then, is to ramp up the usage of nationwide voter files as a means of estimating the effect of election laws on turnout.

# References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2015. "Comparative politics and the synthetic control method." *American Journal of Political Science* 59(2):495–510.

Angrist, Joshua D and Jörn-Steffen Pischke. 2014. *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20(4):437–459.

Ansolabehere, Stephen and Eitan Hersh. 2014. "Voter Registration: The Process and Quality of Lists". In *The Measure of American Elections*, ed. Barry C. Burden and Charles Stewart, III. Cambridge: Cambridge University Press.

Berent, Matthew K, Jon A Krosnick and Arthur Lupia. 2016. "Measuring voter registration and turnout in surveys: Do official government records yield more accurate assessments?" *Public Opinion Quarterly* 80(3):597–621.

Bernstein, Robert, Anita Chadha and Robert Montjoy. 2001. "Overreporting Voting: Why it Happens and Why it Matters." *Public Opinion Quarterly* 65(1):22–44.

Biggers, Daniel R and Michael J Hanmer. 2017. "Understanding the adoption of voter identification laws in the American states." *American Politics Research* 45(4):560–588.

Burden, Barry C. 2000. "Voter turnout and the national election studies." *Political Analysis* 8(4):389–398.

Burden, Barry C. 2018. "Disagreement Over ID Requirements and Minority Voter Turnout." *Journal of Politics* 80(3).

Burden, Barry C., David T. Canon, Kenneth R. Mayer and Donald P. Moynihan. 2014. "Election Laws, Mobilization, and Turnout: The Unanticipated Consequences of Election Reform." *American Journal of Political Science* 58(1):95–109.

Burden, Barry C., David T. Canon, Kenneth R. Mayer and Donald P. Moynihan. 2017. "The Complicated Partisan Effects of State Election Laws." *Political Research Quarterly* 70(3):564–

576.

Calonico, Sebastian, Matias D Cattaneo and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82(6):2295–2326.

Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes. 1960. *The American Voter*. New York: Wiley.

Cooper, Christopher A., Moshe Haspel and H. Gibbs Knotts. 2009. "THe Value of Voterfiles for U.S. State Politics Research." *State Politics and Policy Quarterly* 9(1):102–121.

Dropp, Kyle A. 2013. "Voter ID Laws and Voter Turnout." Working Paper, Draft as of September 21.

Dyck, Joshua J. and James G. Gimpel. 2005. "Distance, Turnout, and the Convenience of Voting." *Social Science Quarterly* 86(3):531–548.

Erikson, Robert S and Lorraine C Minnite. 2009. "Modeling problems in the voter identification—voter turnout debate." *Election Law Journal* 8(2):85–101.

Fraga, Bernard L. 2016*a*. "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout." *American Journal of Political Science* 60(1):97–122.

Fraga, Bernard L. 2016*b*. "Redistricting and the Causal Impact of Race on Voter Turnout." *Journal of Politics* 78(1):19–34.

Fraga, Bernard L., Bradley Spahn and Alan N. Yan. 2017. "Panel Attrition in Voter Files: An Argument for Keeping Dropped Voters." Paper Presented at the Election Sciences, Reform, and Administration Meeting, Portland, OR.

Gimpel, James G., Joshua J. Dyck and Daron R. Shaw. 2004. "Registrants, Voters, and Turnout Variability across Neighborhoods." *Political Behavior* 26(4):343–375.

Green, Donald P. and Alan S. Gerber. 2005. "Preface: Recent Advances in the Science of Voter Mobilization." *Annals of the American Academy of Political and Social Science* 601(September):6–9.

Grimmer, Justin, Eitan Hersh, Marc Meredith, Jonathan Mummolo and Clayton Nall. 2018. "Obstacles to Estimating Voter ID Laws' Effect on Turnout." *The Journal of Politics* 80(3):000–000.

Hajnal, Zoltan, Nazita Lajevardi and Lindsay Nielson. 2017. "Voter identification laws and the suppression of minority votes." *The Journal of Politics* 79(2):363–379.

Hersh, Eitan D. 2015. *Hacking the electorate: How campaigns perceive voters*. Cambridge University Press.

Hersh, Eitan D. and Clayton Nall. 2016. "The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records." *American Journal of Political Science* 60(2):289–303.

Hicks, William D., Seth C. McKee and Daniel A. Smith. 2016. "The Determinants of State Legislator Support for Restrictive Voter ID Laws." *State Politics and Policy Quarterly* 16(4):411–431.

Hicks, William D., Seth C. McKee, Patrick J. Sellers and Daniel A. Smith. 2015. "A Principle or a Strategy? Voter Identification Laws and Partisan Competition in the American States." *Political Research Quarterly* 68(1):18–33.

Highton, Benjamin. 2017. "Voter identification laws and turnout in the United States." *Annual Review of Political Science* 20:149–167.

Holbein, John B. and D. Sunshine Hillygus. 2016. "Making Young Voters: The Impact of Preregistration on Youth Turnout." *American Journal of Political Science* 60(2):364–382.

Hur, Aram and Christopher H. Achen. 2013. "Coding Voter Turnout Responses in the Current Population Survey." *Public Opinion Quarterly* 77(4):985–993.

Imai, Kosuke and Kabir Khanna. 2016. "Improving ecological inference by predicting individual ethnicity from voter registration records." *Political Analysis* 24(2):263–272.

Keele, Luke J and Rocio Titiunik. 2015. "Geographic boundaries as regression discontinuities." *Political Analysis* 23(1):127–155.

Keele, Luke, Rocio Titiunik and José R Zubizarreta. 2015. "Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(1):223–239.

Keele, Luke and William Minozzi. 2013. "How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data." *Political Analysis* 21(2):193–

216.

Khanna, Kabir, Kosuke Imai and Hubert Jin. N.d. "wru: Who are You? Bayesian Prediction of Racial Category Using Surname and Geolocation.".

Leighley, Jan E. and Jonathan Nagler. 2013. *Who Votes Now? Demographics, Issues, Inequality, and Turnout in the United States*. Princeton, NJ: Princeton University Press.

McDonald, Michael P. 2007. "The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics." *Public Opinion Quarterly* 71(4):588–602.

McDonald, Michael P. 2017*a*. "Voter List Information." *United States Elections Project*. `http://http://voterlist.electproject.org`. Last Accessed on April 18, 2017.

McDonald, Michael P. 2017*b*. "Voter Turnout." *United States Elections Project*. `http://www.electproject.org/home/voter-turnout/voter-turnout-data`. Last Accessed on March 7, 2017.

McDonald, Michael P., Enrijeta Shino and Daniel A. Smith. 2015. "Convenience Voting and Turnout: Reassessing the Effects of Election Reforms." Presented at the 2015 New Research on Election Administration and Reform Conference at MIT, Cambridge, MA.

McDonald, Michael P. and Samuel L. Popkin. 2001. "The Myth of the Vanishing Voter." *American Political Science Review* 95(4):963–974.

McKee, Seth C., M.V. III Hood and David Hill. 2012. "Achieving Validation: Barack Obama and Black Turnout in 2008." *State Politics and Policy Quarterly* 12(1):3–22.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Nyhan, Brendan, Christopher Skovron and Rocio Titiunik. 2017. "Differential Registration Bias in Voter File Data: A Sensitivity Analysis Approach." *American Journal of Political Science* 61(3):744–760.

Rosenstone, Steven J. and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: Macmillan.

Sigelman, Lee, Philip W. Roeder, Malcolm E. Jewell and Michael A. Baer. 1985. "Voting and

Nonvoting: A Multi-Election Perspective." *American Journal of Political Science* 29(4):749–765.

Silver, Brian D., Barbara A. Anderson and Paul R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80(2):613–624.

Verba, Sidney, Kay L. Schlozman and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA: Harvard University Press.

Wolfinger, Raymond E. and Steven J. Rosenstone. 1980. *Who Votes?* New Haven, CT: Yale University Press.

# Online Appendix

Figure OA1: Distribution of Turnout Misses Over Time

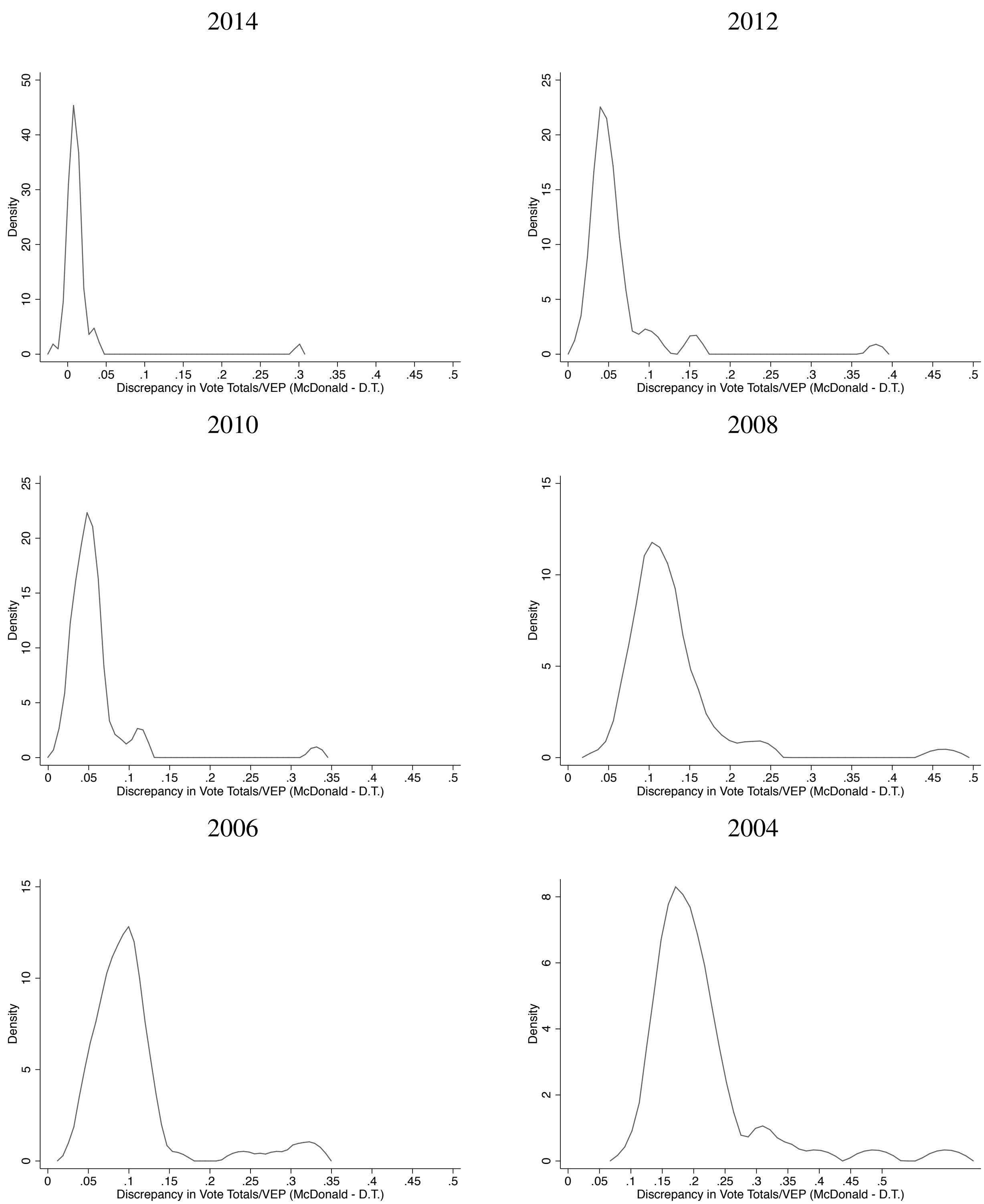


Figure OA2: Registration Date is January 1st by State

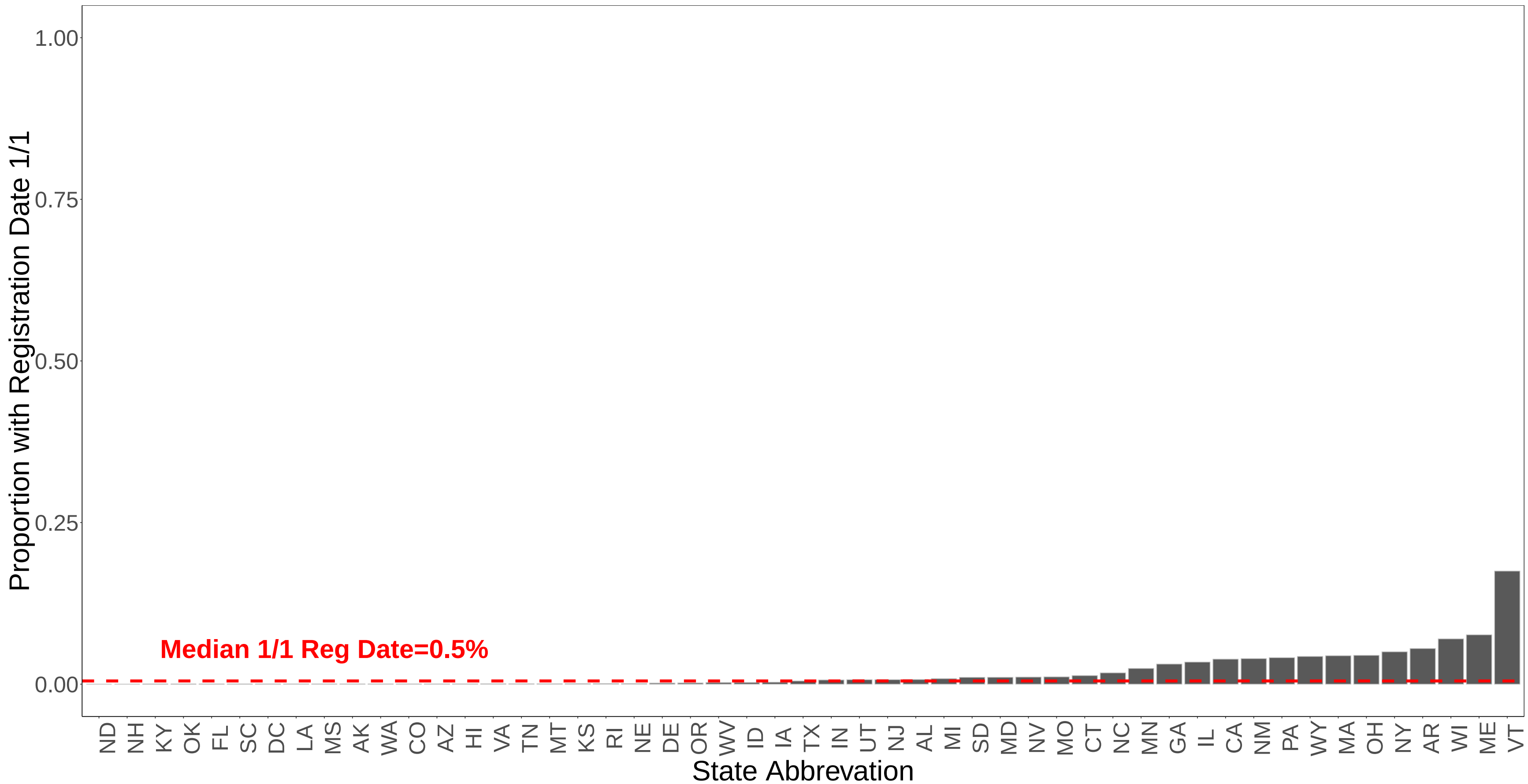


Figure OA3: Difference between McDonald and Data Trust Vote Counts 2014

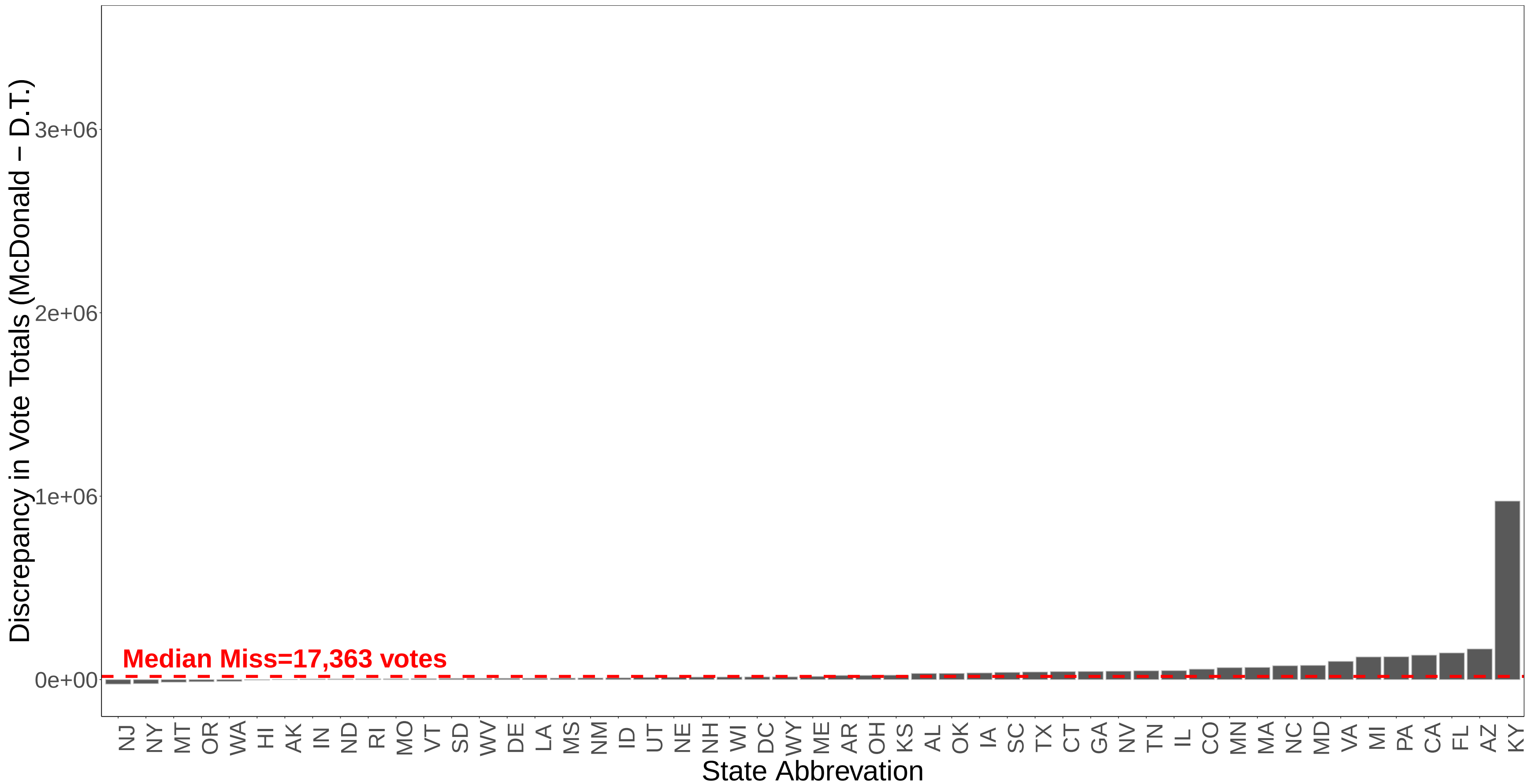


Figure OA4: Difference between McDonald and Data Trust Vote Counts 2012

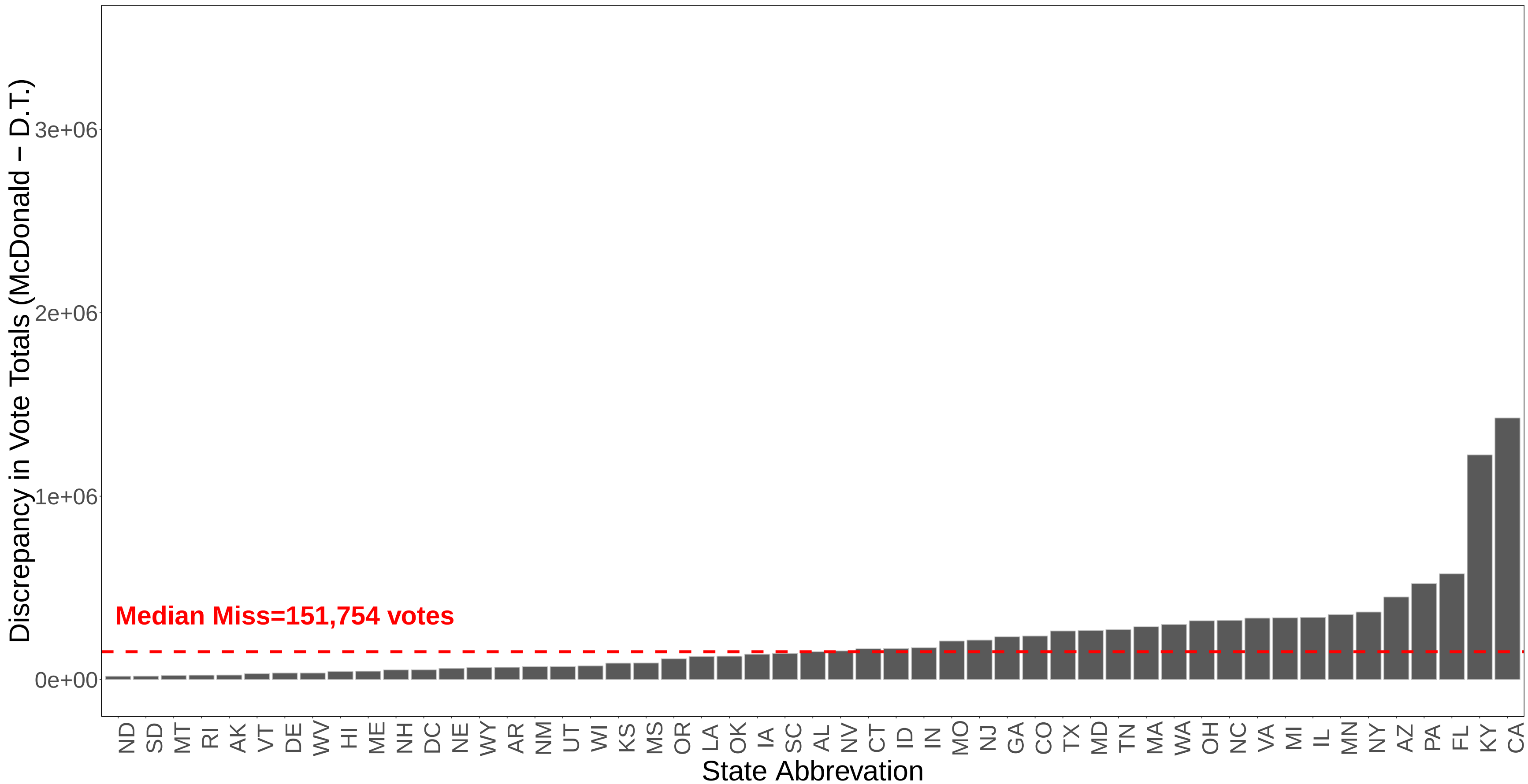


Figure OA5: Difference between McDonald and Data Trust Vote Counts 2010

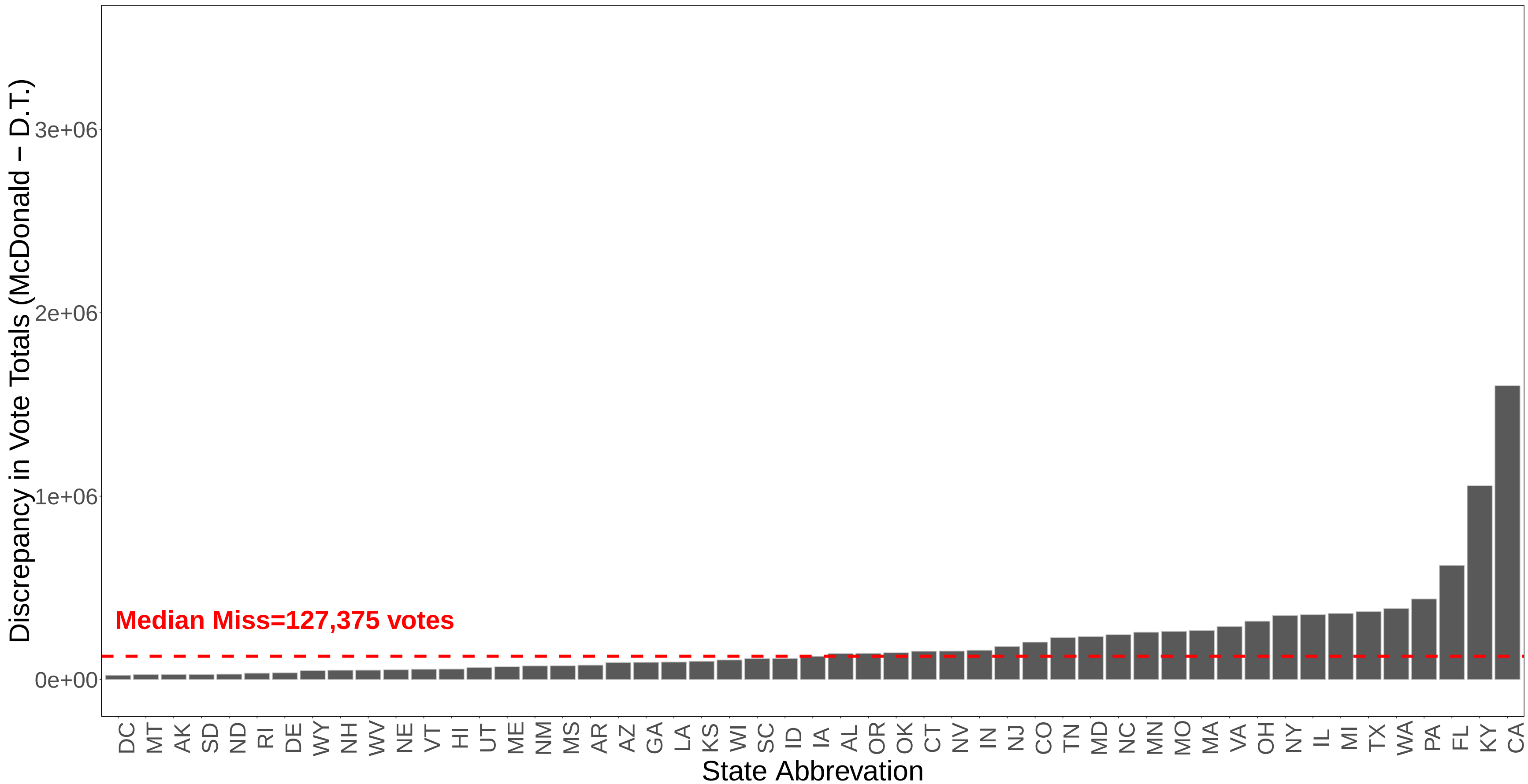


Figure OA6: Difference between McDonald and Data Trust Vote Counts 2008

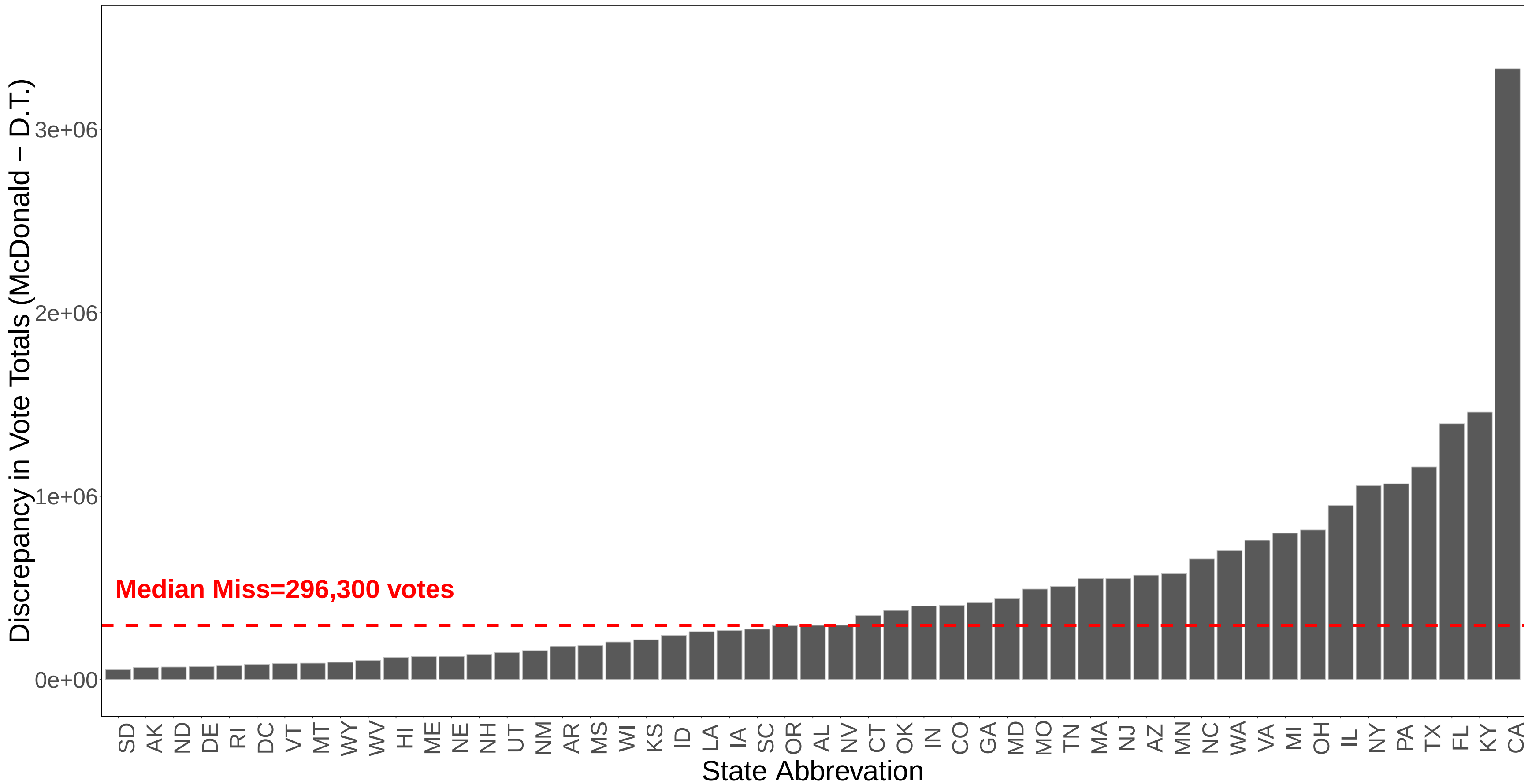


Figure OA7: Difference between McDonald and Data Trust Vote Counts 2006

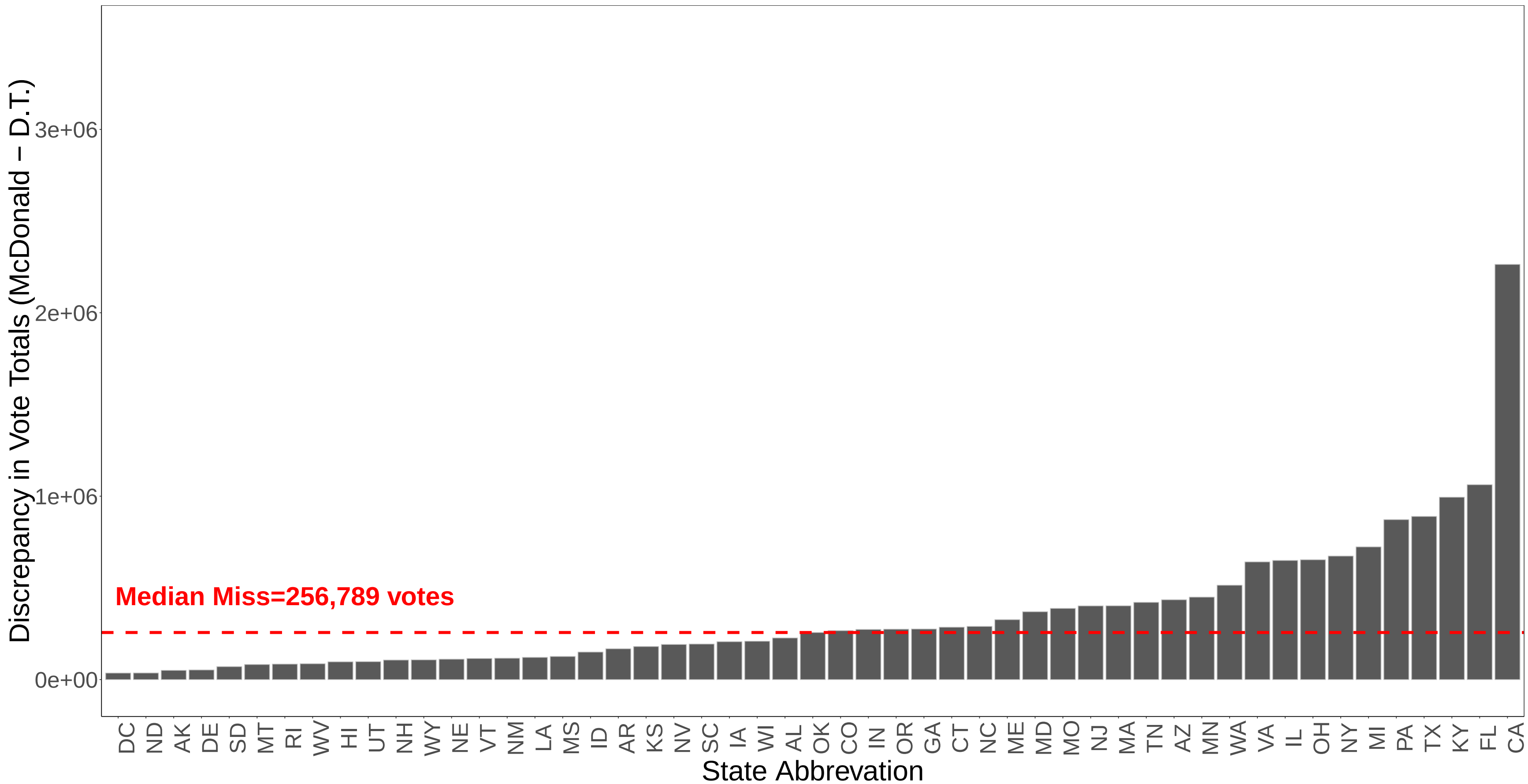


Figure OA8: Difference between McDonald and Data Trust Vote Counts 2004

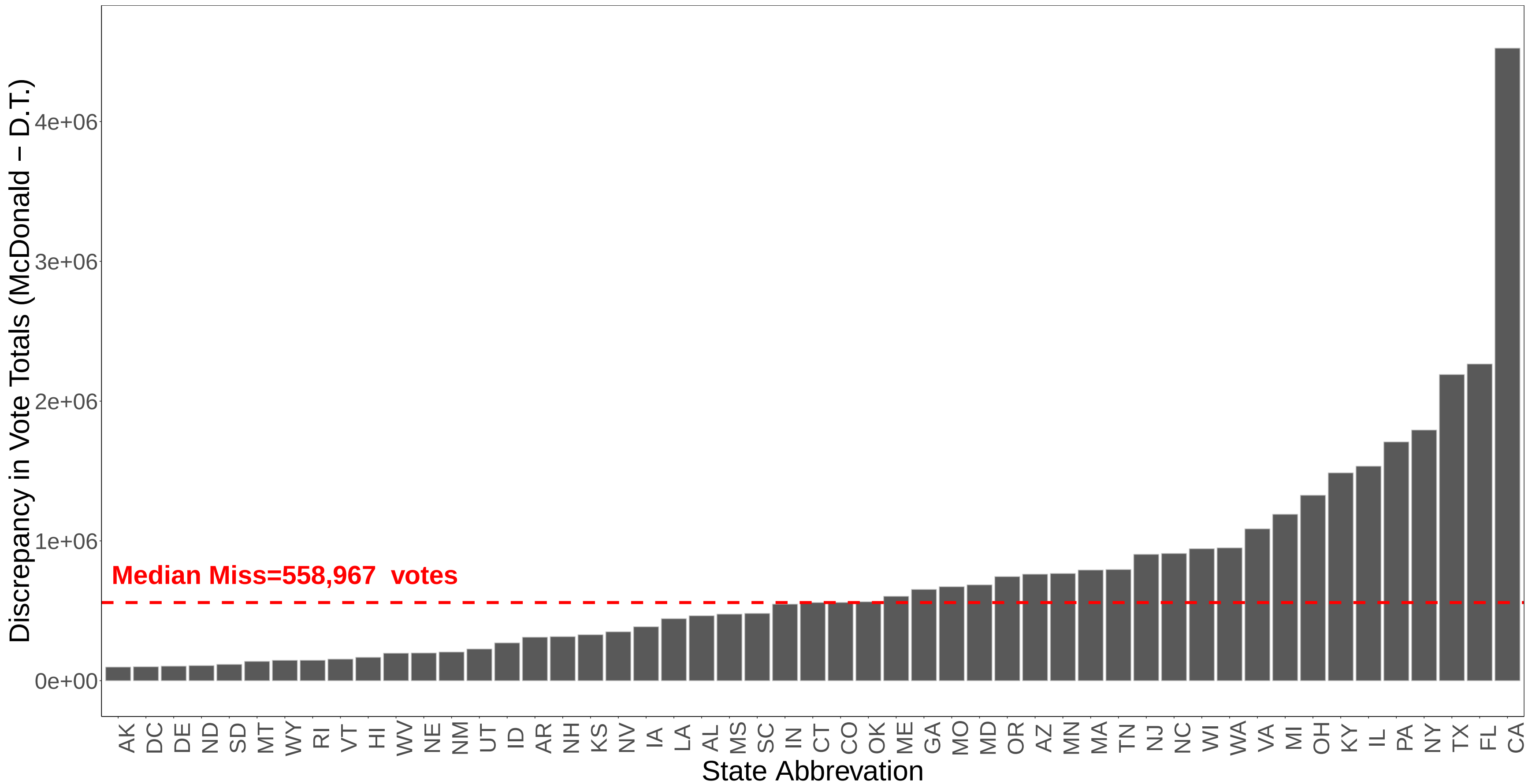


Figure OA9: Difference between McDonald and Data Trust Vote Counts 2002