Validating Self-reported Turnout by Linking Public Opinion Surveys with Administrative Records*

Ted Enamorado† Kosuke Imai‡

First Draft: February 27, 2018
This Draft: March 7, 2018

Abstract

It is widely known that the self-reported turnout rates obtained from public opinion surveys tend to substantially over-estimate the official turnout rate. For example, in recent presidential elections, the self-reported turnout rate from the American National Election Studies (ANES) has consistently exceeded the actual turnout rate by approximately 15 percentage points. However, scholars sharply disagree on what causes the bias of self-reported turnout. Some blame misreporting due to social desirability, others attribute the bias due to non-response or other factors such as the quality of voter records. While in recent years it has become possible to validate self-reported turnout by directly linking surveys with administrative records, most existing validation studies rely on proprietary merging algorithms with limited scientific transparency and yield conflicting results. To shed a light on this debate, we apply a canonical probabilistic record linkage model, implemented via the open-source software package fastLink, to merge two major election studies – the ANES and the Cooperative Congressional Election Survey (CCES) – with a national voter file of over 180 million records. For both ANES and CCES, fastLink successfully produces validated turnout rates close to the actual turnout rates. Using these merged data sets, we find that the bias of self-reported turnout originates primarily from misreporting rather than survey non-response. Our findings suggest that those who are educated and interested in politics are more likely to over-report turnout. Finally, we show that fastLink perform as well as a proprietary algorithm.

*We thank Bruce Willsie of L2 for making the national voter file available and Matt DeBell of ANES and Steffen Weiss of YouGov for technical assistance. We also thank Ben Fifield for his advice and assistance and Steve Ansolabehere for helpful comments.

†Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: tede@princeton.edu, URL: http://www.tedenamorado.com

‡Professor, Department of Politics and Center for Statistics and Machine Learning, Princeton University. Professor of Visiting Status, Graduate Schools of Law and Politics, The University of Tokyo. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: https://imai.princeton.edu

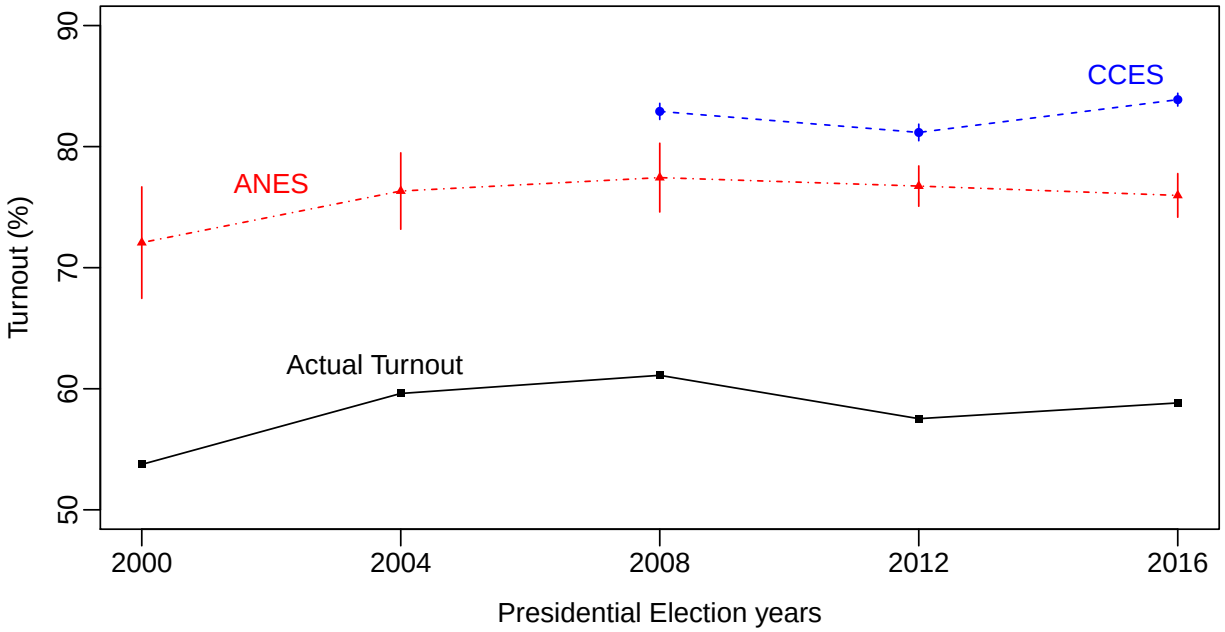


Figure 1: Comparison of Actual and Self-reported Turnout Rates. The actual turnout (solid line with squares) is computed using data from the United States Election Project (http://www.electproject.org), whereas the self-reported turnout rates are based on the American National Election Studies (ANES; dash-dot line with solid triangles for face-to-face interview, dotted line with open triangle for Internet survey) and Cooperative Congressional Election Study (CCES; dashed line with circles), using appropriate survey weights. The vertical bars represent 95% confidence intervals.

1 Introduction

The accuracy of self-reports is essential for ensuring the validity of survey research, and yet many respondents lie or refuse to answer especially when asked survey questions that are sensitive in nature. Social desirability bias and increasingly common unit nonresponse make it difficult to precisely estimate the prevalence of certain attitudes and behavior. A well-known example of this measurement error problem is self-reported turnout rates obtained from public opinion surveys, which have a tendency to substantially over-estimate the actual turnout rate. Figure 1 shows that the gap between self-reported and actual turnout rates has been consistently exceeding 15 percentage points over the last five elections.

The self-reported turnout rates are computed using appropriate survey weights from two major election surveys, the American National Election Studies (ANES) and the Cooperative Congressional Election Study (CCES). The ANES has conducted data collection before and after every presidential election since 1948, whereas the CCES is a large-scale online survey that

has been administered after every election since 2006. While the ANES has used face-to-face interviews, it also conducted an Internet survey in the last three general elections. The actual turnout is obtained from the United States Election Project (McDonald and Popkin, 2001, http://www.electproject.org) and represents the turnout based on the population of eligible voters (see Section 2.1 for more details). The difference between actual and self-reported turnout rates is remarkably consistent during this period. While the actual turnout rate has hovered between 50 and 60 percent, the survey estimates have always stayed above 70 percent with the CCES exceeding 80 percent.

However, scholars sharply disagree on what causes the bias of self-reported turnout rates. Some blame misreporting due to social desirability (e.g., Silver, Anderson and Abramson, 1986; Bernstein, Chadha and Montjoy, 2001), while others attribute the bias due to non-response (e.g., Burden, 2000). Although in earlier years the ANES validated self-reported turnout by manually checking government records, the high cost of this validation procedure led to its discontinuation in the 1990s, making it difficult to resolve the controversy. Fortunately, Congress passed the Help America Vote Act in 2002, mandating that a state develops an official voter registration list. This enabled commercial firms to systematically collect and regularly update nationwide voter registration files (Ansolabehere and Hersh, 2012). Both the ANES and CCES now rely on these commercial firms to validate the self-reported turnout of survey respondents.

Nevertheless, the debate about the causes of the bias of self-reported turnout rates still persists. Most prominently, while Ansolabehere and Hersh (2012) use commercial validation for the 2008 CCES and find that misreporting is the culprit of bias in self-reported turnout, Berent, Krosnick and Lupia (2011, 2016) analyze the 2008 ANES and contend that such findings unreliable because of the poor quality of government records and the errors in matching survey respondents to registered voters in administrative records. In a recent working paper, Jackman and Spahn (2017) validate the self-reported turnout in the 2012 ANES based on commercial validation. They find that over-reporting is responsible for six percentage points whereas non-response bias and inadvertent mobilization effect account for four and three percentage points, respectively. In sum, the existing evidence is mixed as to what biases self-reported turnout in public opinion surveys.

In this paper, we contribute to this literature by examining the validity of self-reported turnout in the 2016 United States presidential election. Our validation study is based on both the ANES

and CCES. We apply a canonical probabilistic model of record linkage, originally proposed by Fellegi and Sunter (1969) and recently improved by Enamorado, Fifield and Imai (2017b), to match survey respondents with registered voters in a nationwide voter file of more than 180 million records. Unlike Ansolabehere and Hersh (2012) and Jackman and Spahn (2017) who rely on a proprietary record linkage algorithm, we use the open-source software package fastLink (Enamorado, Fifield and Imai, 2017a) to maximize the scientific transparency. In addition, unlike Berent, Krosnick and Lupia (2016) who evaluates the performance of deterministic record linkage methods, we consider a probabilistic method that is more commonly used in the statistical literature (e.g., Winkler, 2006; Lahiri and Larsen, 2005). To the best of our knowledge, this paper represents one of the first efforts to examine the empirical performance of a probabilistic record linkage method using large-scale administrative records.

We find that the validated turnout rate for the ANES based on fastLink closely approximates the actual turnout rate when combined with clerical review. For the CCES, the probabilistic record linkage method without clerical review yields the validated turnout rate that is closer to the actual turnout rate. We conjecture that because the CCES is a noisier data set with missing and invalid address entries, clerical review induces false negatives, lowering a validated turnout rate. For both the ANES and CCES, we obtain similar validated turnout rates for pre-election and post-election surveys, suggesting that panel attrition accounts little for the bias in self-reported turnout. We do find, however, that 30 to 40 percent of the matched non-voters falsely report they voted in the election, implying that over-reporting is responsible for much of the bias. This finding is consistent with the conclusion of Ansolabehere and Hersh (2012) but contradict that of Berent, Krosnick and Lupia (2016). Similar to the previous literature, we find that those who are highly educated and interested in politics are more likely to over-report turnout. Finally, using the CCES, we show that the probabilistic record linkage method performs at least as well as the proprietary algorithm.

The rest of the paper is organized as follows. In Section 2, we describe the sampling designs of the ANES and CCES and show that the estimated turnout rates based on these survey data are severely biased. In Section 3, we provide a detailed account of the methods and procedures used to validate the self-reported turnout in these two survey data sets. Section 4 reports empirical findings and Section 5 gives concluding remarks.

2 The 2016 US Presidential Election

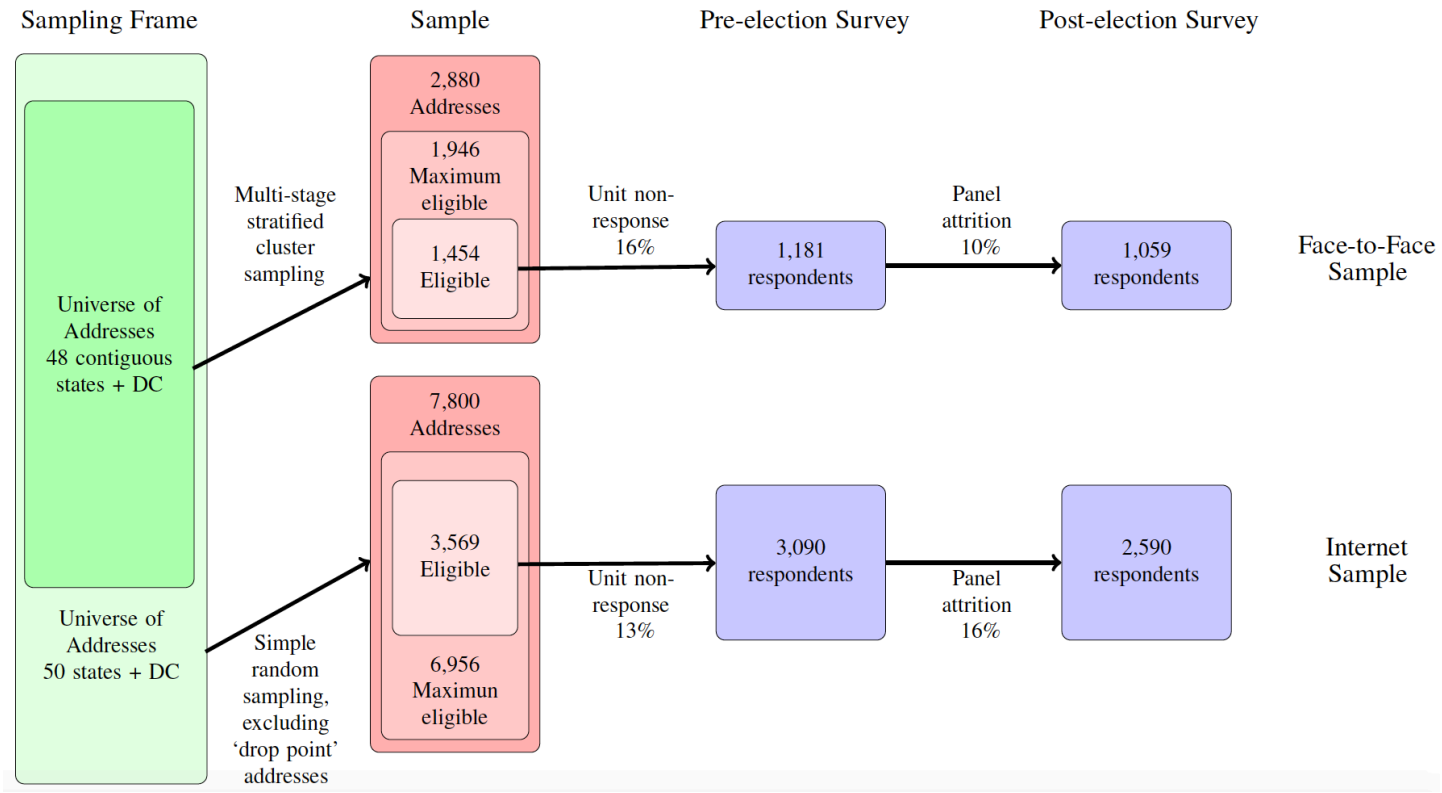
The 2016 US presidential election provides an interesting and important case study for validating the self-reported turnout rates obtained from public opinion surveys. Donald Trump's surprising victory over Hillary Clinton contradicted the prediction of most pre-election polls and as a result raised the question of why polls failed (e.g., American Association for Public Opinion Research, 2017). Researchers have suggested non-response and social desirability biases as possible explanations of polling inaccuracy (e.g., Enns, Lagodny and Schuldt, 2017). As mentioned earlier, these biases may also underlie the gap between self-reported and actual turnout rates, and hence the validation exercise in this particular election should provide useful insights.

In this section, we first summarize the sampling designs of the ANES and CCES in order to characterize the target population of each survey and non-response problems. We also describe the national voter file used in this paper and explain how it relates to the actual turnout rate and the target populations of the two surveys. Finally, we quantify the bias of self-reported turnout rates obtained from the ANES and CCES in the 2016 US presidential election. Along with turnout rates, we also examine the bias of self-reported registration rates when compared to the corresponding rates based on the nationwide voter file.

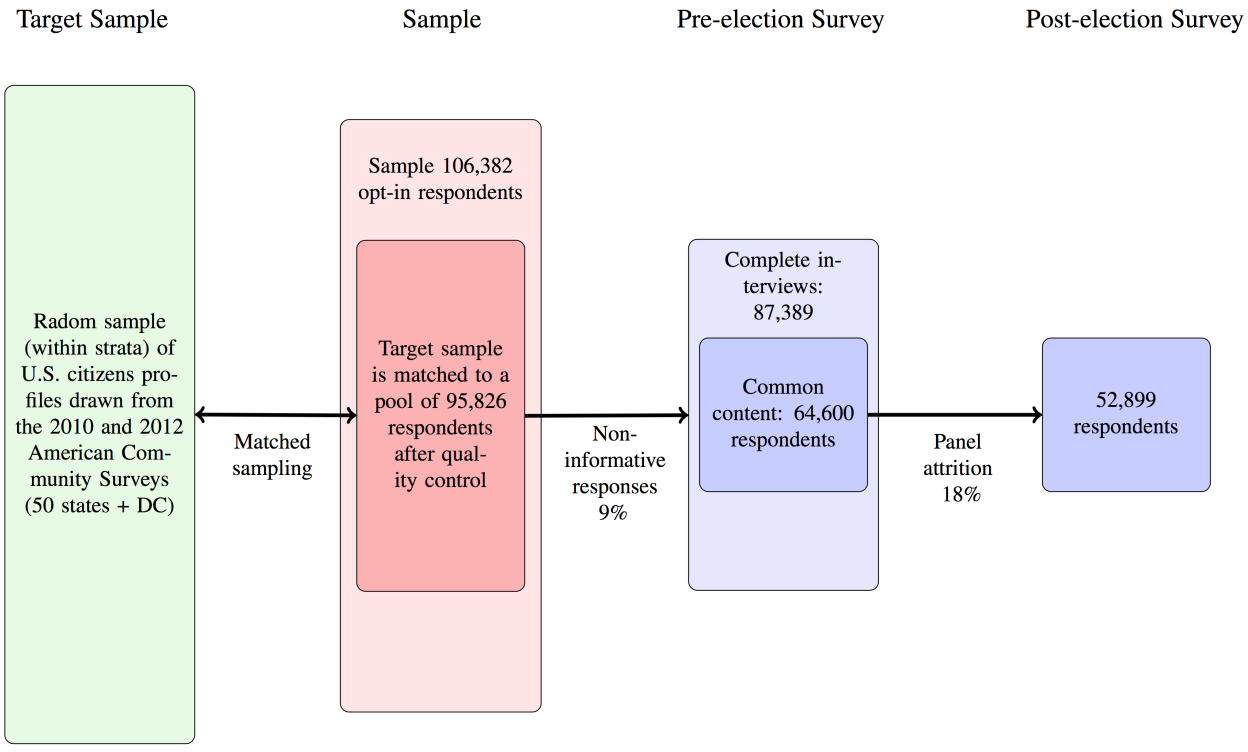
2.1 The Sampling Designs of the ANES and CCES

Figures 2a and 2b schematically summarize the sampling designs of the 2016 ANES and CCES, respectively. For the ANES, there are two modes of interview, face-to-face and the Internet (see American National Election Studies, 2017, for details). As shown in the upper panel of Figure 2a, for the face-to-face sample, the ANES used a multi-stage stratified cluster sampling where 60 counties were randomly selected within each strata defined by regions (excluding Alaska and Hawaii) and other factors. Within a selected county, a certain number of household addresses were chosen at random, yielding a sample of 2880 addresses. Removing some invalid (e.g., non-residential or vacant units, residences occupied by non-citizens) and subsampled out addresses led to 1,946 maximum eligible addresses, from which the final sample of 1,454 eligible addresses are obtained.

Finally, a trained interviewer was sent to each selected household and administered the survey to a randomly selected adult citizen. Thus, the target population is 222.6 million adult US citizens



(a) The Sampling Design of the 2016 ANES



(b) The Sampling Design of the 2016 CCES

Figure 2: The Sampling Designs of the ANES and CCES for the 2016 US Presidential Election. Both surveys have a panel data structure where respondents who answer in the pre-election survey are followed up in the post-election survey.

age 18 or older who reside in contiguous 48 states and D.C. There was a unit non-response rate of 16% in the pre-election survey, whereas the attrition rate from the pre-election to post-election surveys was 10%. The ANES provides sampling weights that are designed to account for this sampling procedure as well as unit non-response, which we will use to estimate the turnout rate for the target population. The nature of the sampling design, however, prohibits us from making inferences about the turnout rate for each state.

Unlike the face-to-face sample, the Internet sample of the ANES was obtained through a simple random sampling of 7,800 addresses from the universe of household addresses in all 50 states and D.C. The bottom panel of Figure 2a shows this sampling design. A letter was mailed to each selected address, and one randomly selected adult citizen was asked to complete an online survey. Following the same criteria as for the face-to-face component, 844 addresses were determined to be invalid. In addition, 3,387 addresses were not part of the study due to unknown eligible status i.e., no reply was received after the letter contact from the ANES. The unit nonresponse rate for the Internet sample was 13%, which is close to that of the face-to-face sample. Although the panel attrition rate was somewhat higher for the Internet sample, reaching 16%, 2,590 respondents completed the post-election survey, which is almost 2.5 times more than the total number of respondents for face-to-face interview. Given this sampling procedure, the target population for the ANES Internet sample is 224.1 million adult citizens age 18 or older who reside in all 50 states and D.C. The ANES provides sampling weights that account for this sampling design and unit non-response. Like the face-to-face sample, survey weights are constructed such that inference should be made at the national rather state level. For a fine-grained account of the sampling design and other methodological details of the data collection efforts of the ANES see DeBell et al. (2016).

The sampling design of the CCES differs from that of the ANES in many ways (see Ansolabehere, Schaffner and Luks, 2017, for details). As summarized in Figure 2b, the CCES first constructed a "target sample" from the respondents who had participated in the 2010 and 2012 American Community surveys where the target population is US adult citizens 18 years or older. Then, the CCES obtained the final sample of potential respondents by selecting, from a pool of opt-in Internet survey respondents, individuals who are matched to the target sample based on the similarity of various respondent characteristics including demographics, party ideology, and polit-

	Self-reported				Administrative		
	ANES			**CCES**	Election	Voter file	
	Overall	Face-to-face	Internet		project	All	Active
Turnout rate (%)	75.96 (0.92)	78.04 (1.80)	75.26 (1.08)	83.79 (0.27)	58.83	57.55	
Registration rate (%)	89.18 (0.71)	89.08 (1.26)	89.22 (0.86)	91.93 (0.21)		80.37	76.57
Target population size (millions of voters)	224.10	222.60	224.10	224.10	232.40	227.60	227.60

Table 1: Comparison of the Estimated Turnout and Registration Rates based on Self-reports and the Administrative Records for the 2016 US Presidential Election. Self-reported turnout and registration rates (with standard errors in parentheses) are obtained from the American National Election Study (ANES) and Cooperative Congressional Election Study (CCES). Since the ANES has two modes of interview, face-to-face and Internet, the estimated turnout and registration rates are computed separately for each mode as well as for the combined sample. The corresponding rates based on administrative records are computed using the voting-eligible population data from United States Election Project and the nationwide voter file from L2, Inc. When using the voter file, we compute the registration rates in two ways, one based on all voters and the other based on active voters only. Each turnout rate has a slightly different target population, which is reflected by the differences in target population size.

ical interests. After removing non-informative responses, a total of 64,600 respondents answered the pre-election survey. With the panel attrition rate similar to that of the ANES Internet sample, the post-election survey had 52,899 respondents. The CCES provides sampling weights, which are designed to balance the characteristics of the final sample with those of the target sample for each state. Thus, the sampling weights enable inference about the target population, which is the same as that of the ANES Internet sample. One major advantage of the CCES is that its large sample size allows for relatively precise estimation of turnout rate for each state.

2.2 The Bias of Self-reported Turnout Rates

We now quantify the bias of self-reported turnout rates obtained from the ANES and CCES by comparing them with the corresponding rates based on administrative records. Along with turnout rates, we also examine self-reported registration rates. Appendix A.1 provides a detailed description of the question wordings and explains how each variable is coded. When computing the bias, it is essential to define the target population of each survey so that the self-reported rates are compared to the actual rates of the same population. The left three columns of Table 1 present

the self-reported turnout and registration rates of the ANES while the fourth column shows the same results for CCES (standard errors that account for survey designs are in parentheses). For the ANES, we present the overall rates as well as the turnout and registration rates separately for the face-to-face and Internet samples. Note that the target population size differs only for the ANES face-to-face sample, which excludes those who reside in Alaska and Hawaii.

We then compare these self-reported rates with the corresponding rates based on the administrative records. We first compute the turnout rate among the voting eligible population (VEP) using the data from the United States Election project (http://www.electproject.org), which utilizes the certified election results and the Census. Since the target populations of ANES and CCES do not exclude individuals on parole or probation, we compute the actual turnout rates as the number of votes for the presidential race divided by the number of eligible voters plus the number of ineligibles minus the total number of prisoners. Unfortunately, we cannot adjust for overseas voters although they are excluded from the target population of both surveys. The reason is that we do not have the information about the number of votes cast by overseas voters alone. As a result, the VEP size has additional 8 to 10 million voters when compared to the target population of the two surveys. Thus, the actual turnout and registration rates presented here should be considered as an approximation to the true rates. As we saw earlier, the gap between self-reported and actual turnout rates is substantial, reaching 17 and 25 percentage points for the ANES and CCES, respectively.

Since our validation procedure involves merging survey data with a nationwide voter file, it is important to examine the differences between the official vote counts and the number of registered voters of our specific voter file who are recorded as casting a ballot for the presidential race. In July 2016, we obtained a nationwide voter file of over 180 million records from L2, Inc., a leading national non-partisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters and consultants for use in campaigns. While by then all states have updated their voter files by including the information about the 2016 election, in the routine data cleaning processes by states and L2, some of the individuals who voted in the election have been removed from voter files because they either have deceased and moved (based on the National Change of Address). As a result, the L2 voter file has a total of 131 million voters who cast their ballots whereas according to the US Election Project,

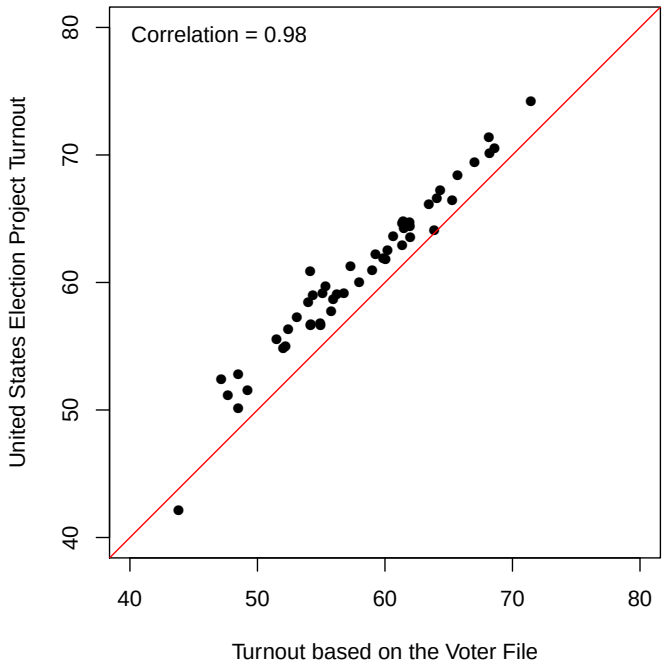


Figure 3: State-level Comparison between the Turnout Rates based on the Voter File and the United States Election Project. The correlation between these turnout rates is high and the average percentage point difference is small.

approximately 136.7 million individuals voted in the election. In addition, the L2 voter file does not contain overseas voters, reducing the total VEP size by about 5 million and the turnout rate by slightly more than one percentage point.

Figure 3 compares state-level turnout rates based on the L2 voter file (horizontal axis) with their corresponding VEP turnout rates from the US Election Project (vertical axis). Recall that deceased voters and those who moved across states have been removed from the voter file, whereas they are included in the VEP turnout calculation. As expected, we find that the turnout rate based on the voter file is lower than the actual turnout. The median difference is 2.7 percentage points whereas the standard deviation is one percentage point. However, the correlation between the two is remarkably strong, reaching 0.98. Even using the turnout rate based on the voter file, the biases of self-reported turnout rates are substantial.

We can also compute the registration rate using the voter file. Since the voter file lists everyone who is registered to vote, we can divide the total number of records in the voter file by the target population size. In the voter file, we have approximately 182 million records among a total of 8.6 million records which are classified by some states as "inactive voters." The definition of inactive voters differs from one state to another (and some states do not have such classification), but

these voters did not turn out in a certain number of consecutive elections and states were unable to contact them. After being placed on the inactive voter list for a few years, these records will be purged entirely by states. Typically, if inactive voters show up to vote at a polling station on an election day, they would have to provide a proof of residence. This suggests that inactive voters may answer in a survey they are not registered. Therefore, we compute the registration rate in two ways, one based on all records in the voter file and the other based on active voters alone. Similar to the self-reported turnout rates, the self-reported registration rates are much greater than those based on the voter file. The gap is about 10 percentage points if we use all records, whereas it is closer to 15 percentage points when compared to the registration rate based on active voters alone.

Finally, it is interesting to note that the magnitude of bias is much greater for these two election studies than the Voter Supplement of the Current Population Survey (CPS), which is an additional questionnaire of the CPS focusing on voting and registration. Historically, the CPS has consistently produced self-reported turnout estimates that are closer to the actual turnout rates than the ANES. For example, for the past three general elections, the difference between the CPS self-reported turnout estimate and the voting-eligible population turnout reported by the Election Project has been of at most of three percentage points. Recently, some scholars have pointed out that the CPS treats those who dropped out or refused to answer the turnout question (Hur and Achen, 2013). However, this does not appear to explain the smaller magnitude of bias as the nationwide turnout rate without this coding scheme is 61.38% ($s.e.$ = 1.48). We leave to future research the question of why the CPS yields more accurate self-reported turnout rates than the ANES and CCES.

3 Linking Surveys with Administrative Records

In this section, we describe how we linked the ANES and CCES with the national voter file, using a canonical probabilistic record linkage model. Through the research collaboration agreements with the ANES and YouGov, we obtained access to the de-anonymized information for each one of the 4,271 respondents (1,181 and 3,090 for the face-to-face and Internet samples, respectively) that were part of the 2016 ANES as well as 64,600 respondents that were part of the 2016 CCES. We then use this information to link the survey data with the voter file.

3.1 Preprocessing Names and Addresses

As emphasized by Winkler (1995), a key step for a successful merge is to standardize the fields that will be used to link two datasets. Accordingly, we made every effort to parse the names and addresses used in the ANES and CCES uniformly so that their formats match with those of the corresponding fields in the nationwide voter file. For example, the full name of an individual is divided into the first, middle, and last names, while the address is parsed into house number, street name, zip code, and apartment number. As shown below, the names and addresses recorded in the ANES are less noisier than those of the CCES.

The ANES makes use of data from the United States Postal Service to ensure that the invitation letter can be delivered to the sampled addresses. As a result, the ANES address data are of high quality. In contrast, the respondent names are self-reported and each name is represented by a string, which we parsed into the first, middle, and last names. For self-reported registered voters, whenever available, we use the name, which they said they had used for registration (3,623 records or 85%). If no name was provided (either because an individual reported not having registered to vote or failed to provide a name), we use the name on a check sent as monetary compensation for their participation in the survey (464 records or 11%). For the remaining respondents, we use the names of individuals whom the ANES intended to interview (184 records or 4%).

In the case of the CCES, both addresses and names are self-reported. Consequently, we manually parsed each name and address for all of the 64,600 respondents in order to make their format comparable to that of names and addresses in the nationwide voter file. In the case of names, we followed a similar strategy as the one used for the ANES. That is, we divided a name string into three components: first, middle, and last names. However, the names of almost three percent of respondents (1,748 individuals) were missing.

As noted above, the CCES respondents self-report their address as well, and each of those addresses was stored as a single string variable. We first used `preprocText()` function in fastLink to standardize each address according to the USPS Postal Address Information System (see `https://pe.usps.com/cpim/ftp/pubs/pub28/pub28.pdf` for more information). This follows the same procedure used by the ANES to clean their sample of addresses. We then divided a standardized address into house number, street name, zip code, and apartment number. Unlike the ANES, which has no missing value, more than seven thousand records (or 11 percent) of the

	ANES						CCES	
	All		Face-to-face		Internet			
	Cases	%	Cases	%	Cases	%	Cases	%
Names								
Missing value (first or last name)	66	1.55	53	4.49	13	0.42	1,748	2.71
Initials for first and last name	0	0.00	0	0.00	0	0.00	3,274	5.07
Initials for first name but last name complete	16	0.38	7	0.01	9	0.29	506	0.78
Complete name	4,189	98.07	1,129	95.60	3,068	99.29	59,072	91.44
Addresses								
Missing value	0	0.00	0	0.00	0	0.00	7,465	11.55
P.O. Box	0	0.00	0	0.00	0	0.00	1,665	2.58
Complete address	4,271	100.00	1,181	100.00	3,090	100.00	55,470	85.87
Number of respondents	4,271		1,181		3,090		64,600	

Table 2: The Data Quality of the Name and Address Fields for the 2016 ANES and 2016 CCES.

CCES respondents did not report their addresses.

Table 2 summarizes the results of preprocessing names and addresses. We find that the percentage of complete names across surveys is quite high, exceeding 90% for both surveys. Note that the ANES has a higher proportion of complete names, regardless of its interview mode, than the CCES, which has some cases of missing names and uses of initials. However, there is an important difference in address fields between the two surveys. Since the ANES adopts the sampling design based on the list of residential addresses, all addresses are complete. In contrast, the CCES relies on the self-reported addresses by its respondents, resulting in the non-response rate of over 10% and some use of P.O. Box. Indeed, the CCES has 8,716 cases (13.5% of the pre-election sample) without any information about names or a valid residential address. This makes it more challenging to merge the CCES data with the voter file.

3.2 Merge Procedure

Having standardized the linkage fields, we separately merge the ANES and CCES with the nationwide voter file. Since the nationwide voter file contains more than 180 million records, merging a survey data set with the voter file all at once would result in a total of over 756 billion and 18

trillion comparisons for the ANES and CCES, respectively. Therefore, we first subset the survey and voter file data into 102 blocks, defined by state of residence (50 states plus Washington DC) and gender (male and female). Once the within-state merge is done for each block, we then conduct the across-state merge focusing on survey respondents who are not matched with registered voters through the within-state merge.

In the case of the ANES, the block size ranges from 48,315 pairs (Hawaii/Female: ANES = 3, Voter file = 16,105) to 705 million pairs (California/Female: ANES = 225, Voter file = 3,137,276) with the median value of 11 million pairs (Idaho/Male: ANES = 28, Voter file = 426,636). For the CCES, the block size ranges from more than 3 million (Wyoming/Male: CCES = 45, Voter file = 88,849) to 25 billion pairs (California/Male: CCES = 3,073, Voter file = 8,326,559) with the median value of 301 million pairs (Iowa/Female: CCES = 394, Voter file = 764,169).

Within each block, we conducted the data merge using the following variables: first name, last name, age, house number, street name, and zip code. We applied a canonical probabilistic record linkage model, which was originally proposed by Fellegi and Sunter (1969). Enamorado, Fifield and Imai (2017b) improved the implementation of the algorithm used to fit this model so that it is possible to merge large scale data sets with millions of records. Throughout the merge process, we use the open-source package fastLink (Enamorado, Fifield and Imai, 2017a) to fit the model to our data so that the procedure is transparent.

The model is fit to the data based on the agreement patterns of each linkage field across all possible pairs of records between the two data sets $\mathcal{A}$ and $\mathcal{B}$. We use three levels of agreement for the string valued variables (first name, last name, and street name) based on the Jaro-Winkler distance with 0.85 and 0.94 as the thresholds (see e.g., Winkler, 1990). We also use three levels of agreement for age based on the absolute distance between values, with 1 and 2.5 years as the thresholds used to separate agreements, partial agreements, and disagreements (see ANES (2016) for a similar choice). For the remaining variables (i.e., house number and postal code), we utilize a binary comparison indicating whether they have an identical value.

Formally, if we use a binary comparison for variable k, we define $\gamma_k(i, j)$ to be a binary variable, which is equal to 1 if record i in the data set $\mathcal{A}$ has the same value as record j in the data set $\mathcal{B}$. If the variable uses a three-level comparison, then we define $\gamma_k(i, j)$ to be a factor variable with three levels, in which 0, 1, and 2 indicate that the values of two records for this variable are

different, similar, and identical, respectively. Based on this definition, the record linkage model of Fellegi and Sunter (1969) can be written as the following two-class mixture model with the latent variable M_{ij}, indicating a match $M_{ij} = 1$ or a non-match $M_{ij} = 0$ for the pair (i, j),

$$\gamma_k(i,j) \mid M_{ij} = m \overset{\text{indep.}}{\sim} \text{Discrete}(\boldsymbol{\pi}_{km}) \tag{1}$$

$$M_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda) \tag{2}$$

where $\boldsymbol{\pi}_{km}$ is a vector of length L_k, which is the number of possible values taken by $\gamma_k(i,j)$, containing the probability of each agreement level for the kth variable given that the pair is a match ($m = 1$) or a non-match ($m = 0$), and λ represents the probability of match across all pairwise comparisons. The model assumes (1) independence across pairs, (2) independence across linkage fields conditional on the latent variable M_{ij}, and (3) missing at random conditional on M_{ij} (Enamorado, Fifield and Imai, 2017b). These assumptions considerably simplify the computation, enabling researchers to handle large data sets.

Once the model is fitted to the data, we compute the posterior probability of match using the Bayes rule based on the maximum likelihood estimates of the model parameters,

$$\begin{aligned} \xi_{ij} &= \Pr(M_{ij} = 1 \mid \boldsymbol{\delta}(i,j), \boldsymbol{\gamma}(i,j)) \\ &= \frac{\lambda \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k - 1} \pi_{k1\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)}}{\sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k - 1} \pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)}} \end{aligned} \tag{3}$$

where $\delta_k(i,j)$ indicates whether the value of variable k is missing for pair (i,j) (a missing value occurs if at least one record for the pair is missing the value for the variable).

We say that record j is a potential match of record i if the posterior match probability ξ_{ij} is the largest among all pairs that involve record i. Formally, define the following maximum posterior match probability for record i as follows,

$$\zeta_i = \max_{j \neq i} \xi_{ij} \tag{4}$$

If there are more than one record whose posterior match probability is equal to ζ_i, then we randomly select one of them as a match. Fortunately, in the current applications, there was no

tie when ζ_i is reasonably high, e.g., $\zeta_i \geq 0.75$, and hence random sampling has little effect. This procedure yields one-to-one match for each respondent i with the posterior matching probability of ζ_i.

An important concern with our blocking strategy is that we may fail to match an individual whose residential address has changed between the day of survey interview and the date when our voter file was updated. It is also possible that people were registered to vote in a residential address different from the address they reported in the surveys. To identify these individuals, we take all survey respondents whose posterior match probability ζ_i is less than 0.75 and then merge them with registered voters in other states. There were a total of 1,100 such respondents for the ANES and 22,711 respondents for the CCES.

To conduct this across-state merge, we first subset the nationwide voter file such that it only contains the registered voters whose names are close to the remaining survey respondents. As before, we use the Jaro-Winkler string distance of 0.92 or above as the threshold. This reduces the number of registered voters from over 180 million to 14 million. Using fastLink, we then find, for each survey respondent, a registered voter who has the same name (first, middle, and last) and the identical age where the names with the Jaro-Winkler distance of 0.94 or above are coded as same. This yields 51 and 874 additional matches for the ANES and CCES, respectively, and for these matches the posterior match probability is close to 1. For those respondents who are not matched, we use the matches from the within-state merge.

As an optional final step, we conduct a clerical review of each respondent, which is recommended by some in the literature (e.g., Winkler, 1995), and set the posterior match probability to zero for those respondents who, our clerical review suggests, do not have a valid match. We caution that a clerical review may not be useful when the data contain many missing or mismeasured variables. In such cases, a clerical review may increase false negatives while reducing false positives. In our applications, as shown in Section 3.1, the names and addresses are more complete for the ANES than for the CCES. As a result, a clerical review may be more appropriate for the ANES.

Our clerical review discards 280 and 4,115 matches for the ANES and CCES, respectively. For example, 124 cases in the ANES and 2,335 in the CCES are removed because we suspect that a survey respondent is matched with a different individual in the same household who has the

		Pre-election		Post-election		Registration rate		
			clerical		clerical	Voter file		
		fastLink	review	fastLink	review	all	active	CPS
	Overall	76.54	68.79	77.15	69.85	80.37	76.57	70.34
		(0.63)	(0.71)	(0.67)	(0.76)			(1.40)
ANES	Face-to-face	75.32	67.82	75.64	69.12	80.22	76.43	70.40
		(1.21)	(1.36)	(1.27)	(1.42)			(1.39)
	Internet	77.00	69.16	77.77	70.15	80.37	76.57	70.34
		(0.74)	(0.83)	(0.79)	(0.90)			(1.40)
CCES		66.60	58.59	70.52	63.57	80.37	76.57	70.34
		(0.18)	(0.19)	(0.19)	(0.21)			(1.40)

Table 3: Match Rates from the Results of Merging the ANES and CCES with the Nationwide Voter File. For the ANES, we compute the match rates separately for the face-to-face and Internet samples as well as together for the overall sample. Merging is based on the probabilistic model alone ("fastLink") and the model plus clerical review ("clerical review"). Standard errors are given within parentheses. For the sake of comparison, we also present the estimated registration rates from the voter files (all registered voters "all" and active voters only "active") as well as the self-reported registration rate from the Current Population Survey (CPS). Each validated turnout rate is computed for the target population of corresponding survey estimate.

same name but an age difference of more than 5 years, suggesting that the individual is likely to be a parent or child with the same name. Similarly, we discard 39 cases in the ANES and 59 in the CCES, where matched individuals have the same name and age, but a different address and middle name. Finally, we remove 60 cases in the ANES and 1,404 in the CCES where individuals had the same address and age, but the names were different.

3.3 Match Rates

To summarize the results of merge, we use the match rate defined as, $\sum_{i=1}^{N} \zeta_i / N$ where N is the total number of survey respondents. Table 3 presents the match rates for the ANES and CCES using the pre-election and post-election survey respondents. For the ANES, we present the match rate separately for the face-to-face and Internet samples as well as for the combined sample ("Overall"). The results are based on the probabilistic model alone ("fastLink") and the model plus clerical review ("clerical review").

For the sake of comparison, we also present the two estimates of registration rate based on the voter file for the target populations for surveys. The first ("all") is the total number of voters in

the voter file divided by the number of eligible voters. However, these registration rates are likely to overestimate the true rates because not all records of the voter file are valid: some voters may have deceased or moved. For this reason, as explained earlier, in some (but not all) states, the Secretary of State office labels voters "inactive" before purging them from the voter file, if, for example, they do not turn out for multiple elections and they cannot be reached by mail. The second estimate ("active") uses the total number of active voters as the numerator. Since the exact definition of active voters varies by states and some states do not distinguish active and inactive voters, these estimates may not approximate the actual registration rate. It is possible that survey respondents may think they are registered even though they are classified as inactive voters or even removed from the voter file. In the final column, we also present the estimated registration rate based on self-reports from the CPS.

For the ANES, the match rates based on the probabilistic model alone ("fastLink") are similar to the registration rates based on active voters. After the clerical review, however, the estimates become closer to the self-reported registration rates from the CPS. There is little difference in results between the pre-election and post-election samples as well as between the interview mode. For the CCES, the match rates are generally lower than those of the ANES. This makes sense since the CCES contains a larger number of missing and misreporting entries for names and addresses. This suggests that for the noisy data like the CCES, probabilistic models alone might perform better because clerical review may end up with a greater number of false non-matches while reducing false positives. Finally, for the CCES, the match rate for the pre-election sample is about four to five percentage points lower than those for the post-election sample. This suggests that unlike the ANES, the weighting adjustment may not be sufficient to adjust for attrition in the CCES.

Merging the 2008 ANES respondents with the voter files for six states, Berent, Krosnick and Lupia (2016) find that the match rates are significantly lower than the registration rates. The authors use this as the evidence to argue that the validated turnout rates are lower than self-reported turnout rates not because survey respondents misreport but because merging methods fail to match some respondents who voted with voter registration records. We find a similar pattern: the match rates based on our probabilistic approach are generally lower than the registration rates based on the voter file. However, as explained above, the registration rates based on the voter

		Pre-election		Post-election		Actual turnout	
		fastLink	clerical review	fastLink	clerical review	Voter file	Election project
ANES	Overall	63.59 (0.91)	58.09 (0.93)	64.97 (0.96)	59.78 (1.00)	57.55	58.83
	Internet	62.59 (1.06)	57.04 (1.08)	63.99 (1.15)	58.55 (1.18)	57.55	58.83
	Face-to-face	66.46 (1.76)	61.12 (1.78)	67.62 (1.68)	63.10 (1.83)	57.58	58.86
CCES		54.11 (0.31)	48.50 (0.31)	55.67 (0.37)	50.25 (0.37)	57.55	58.83

Table 4: Validated Turnout Rates among the Survey Respondents from the 2016 ANES and CCES. The validated turnout rates obtained from the probabilistic model alone ("fastLink") and the model plus clerical review ("clerical review") are compared to the actual turnout rate for the corresponding target population based on the voter file and the data from the United States election project. The standard errors are given in parentheses.

file are likely to overestimate the true rates because of inactive voters who remain in the voter file. Thus, our interpretation of this result differs from that of Berent, Krosnick and Lupia (2016). Below, we present evidence that over-reporting is primarily responsible for the bias in self-reported turnout.

4 Empirical Findings

In this section, we present the results of our turnout validation. We begin by showing validated turnout rates and then examine the potential sources of bias in self-reported turnout rates. Finally, we identify the types of voters who tend to misreport their turnout and compare our validation results with those of a commercial vendor.

4.1 Validated Turnout Rates

To obtain the validated turnout rate, we compute the weighted average of the binary turnout variable among matched voters in the voter file where the posterior match probability ζ_i is used as the (unnormalized) weight. Table 4 presents the validated turnout rates among the survey respondents from the pre-election and post-election surveys of the 2016 ANES and CCES. As in Table 3, we compare the results obtained from the probabilistic model alone ("fastLink") and the

model plus clerical review ("clerical review") with actual turnout rates based on the voter file ("Voter file") and the United States election project ("Election project"). The standard errors that account for sampling design and unit non-response, are given in parentheses.

Our main findings about turnout rates are consistent with those about registration rates given in Table 3. For the ANES, we find that the validated turnout rates directly obtained from fastLink are at least five percentage points greater than the actual turnout rates. However, clerical review helps close this gap, yielding the validated turnout rates that are within the sampling error of the actual turnout rates. For the sample of face-to-face interview, the validated turnout rates are higher than the Internet sample though the standard errors are greater. This bias for the face-to-face sample appears to be driven largely by the sampling weights as the unweighted validated turnout rates are similar to those of the Internet sample and hence closer to the actual turnout rates. For example, the unweighted post-election validated turnout after clerical review is 59.12% ($s.e. = 1.55$) (see Table 8). Appendix A.2 provides more information about the sampling weights.

For the CCES, the validated turnout rates directly obtained from fastLink are closer to the actual turnout rates than those based on the model and clerical review. The reason for this difference is the same as the one discussed earlier. Because the CCES contains many misreported and missing entries especially for addresses, clerical review ends up removing the potential matches involving these records and hence introducing false negatives. This suggests that clerical review may be ineffective for noisy data. We also note that the validated turnout rates based on the model and clerical review are similar to the result obtained by YouGov based on a voter file provided by a commercial firm, Catalist. In Section 4.4, we further compare our validated turnout with that based of data from Catalist.

4.2 Possible Sources of Bias in Self-reported Turnout

What are the possible sources of differences between self-reported and validated turnout rates? The literature suggests misreporting, attrition, and mobilization as the main culprits. Below, we examine each potential cause in turn (the results on mobilization are forthcoming) and show that misreporting accounts for much of the bias of self-reported turnout.

			Registered		Post-election
		Not registered	Did not Vote	Voted	attrition
ANES	fastLink	8.11	14.45	81.74	55.66
		(1.58)	(1.74)	(0.86)	(2.41)
	Clerical review	0.90	5.97	77.44	48.27
		(0.78)	(1.21)	(0.99)	(2.41)
	Number of respondents	342	444	2,862	622
CCES	fastLink	16.37	10.15	73.05	24.02
		(0.84)	(0.73)	(0.28)	(0.60)
	Clerical review	8.04	4.67	68.66	16.44
		(0.73)	(0.59)	(0.30)	(0.51)
	Number of respondents	4,684	3,237	44,796	11,701

Table 5: Validated Turnout Rates among Survey Respondents with Different Responses to the Turnout Questions in the ANES and CCES. "Post-election attrition" refers to the group of survey respondents who did not answer the turnout questions due to attrition. Standard errors that account for the sampling designs and unit non-response are given within parentheses.

4.2.1 Misreporting

We first consider misreporting as a potential source of bias in self-reported turnout. Table 5 presents the validated turnout rates among survey respondents with different responses to the turnout questions of the ANES and CCES. We find that about 20% of the ANES respondents who said they had voted in the post-election survey did not turn out according to the voter file, whereas the corresponding estimated proportion of misreporting for the CCES is about 30%. Compared to the probabilistic model alone ("fastLink"), the use of clerical review ("clerical review") increases the estimated misreporting rate by several percentage points for both surveys. Because a majority of respondents said they had voted (78% for the ANES and 85% for the CCES), over-reporting is mostly responsible for the upward bias in self-reported turnout.

In contrast, the results from fastLink shows that approximately 15% (10%) of the ANES (CCES) respondents who said they had registered but had not voted were matched with registered voters who had voted in the 2016 election. When clerical review is conducted after fastLink is used, this proportion is reduced to 6% (4%). In addition, less than 10% (15%) of the ANES (CCES) respondents who said they had not registered actually turned out in the election according to the voter file. Again, clerical review reduces this number to less than 1% for the ANES and 6% for the

		Voters		Non-voters		
		%	Cases	%	Cases	Total
ANES	fastLink	95.68	2,436	33.66	378	2,814
		(0.50)		(3.01)		
	Clerical review	98.50	2,258	30.84	290	2,548
		(0.32)		(3.48)		
CCES	fastLink	92.70	33,329	43.49	3,901	37,230
		(0.36)		(1.25)		
	Clerical review	96.33	30,741	44.35	2,836	33,577
		(0.32)		(1.75)		

Table 6: Self-Reported Turnout Rates among Matched Voters and Non-voters. In the "Voters" ("Non-voters") column, we present the self-reported turnout rate among the survey respondents who are validated to have voted (have abstained) in the 2016 election. More than 30% (40%) of the ANES (CCES) survey responded who did not vote reported they had voted. Standard errors are given within parentheses.

CCES. These discrepancies, while smaller, represent potential misreporting that may contribute to an downward bias. However, since the fewer number of survey respondents said they had not voted, this potential under-reporting contributes little to the bias of the overall self-reported turnout rates.

Table 6 provides additional evidence that survey respondents tend to over-report turnout. In this table, we present the self-reported turnout rates among the survey respondents whom we were able to match with registered voters in the voter file. For the results based on fastLink without clerical review, we use the posterior match probability in equation (4) to weight each observation. We find that although most of those who were validated to have voted did not misreport, more than 30% (40%) of the ANES (CCES) survey responded who did not vote reported they had voted. This finding is consistent with that of Ansolabehere and Hersh (2012). Although matched non-voters may differ from non-voters who are not matched, our finding suggests that the unmatched non-voters may also over-report their turnout, leading to a substantial over-reporting. Our finding contradicts the claim put forth by Berent, Krosnick and Lupia (2016) that survey respondents do not misreport turnout. These authors show that matched respondents tend not to lie. However, they do not separate matched voters from matched non-voters, and as a result may have overlooked the tendency of matched non-voters to over-report.

4.2.2 Attrition

Next, we examine the consequences of attrition. The last column of Table 5 presents the validated turnout among those who dropped out after the pre-election survey and did not answer the post-election survey. We observe that the validated turnout rate for the ANES dropouts is similar to the overall turnout, suggesting that attrition does not substantially bias the results. For the CCES, those who did not answer the post-election survey have a much lower validated turnout rate, implying that attrition may have contributed to the bias of self-reported turnout.

This pattern is consistent with Table 4, which shows the similarity of the validated turnout rates between the pre-election and post-election surveys for the ANES, but not for the CCES. In contrast with some previous work in the literature (e.g., Burden, 2000), this finding suggests that attrition is unlikely to explain the gap between the self-reported and actual turnout rates for the ANES though it may be responsible for some, but not all, of the bias for the CCES. Sampling weights of the ANES appear to be able to properly adjust for the possible bias due to unit and item non-response.

4.3 Who Over-reports Turnout?

To determine who over-reports, we conduct a regression analysis using the sample of validated non-voters alone. The outcome variable is binary and equals one if a respondent self-reported that she voted but our turnout validation based on fastLink and clerical review found that she did not. In our weighted logistic regression model with survey weights, we include several covariates used in the literature (e.g., Ansolabehere and Hersh, 2012): age, squared age, marital status, highest level of educational attainment, gender, race, income, partisanship, religiosity, and ideology. Appendix A.5 explains the coding rules we use to harmonize covariates across the two surveys to facilitate the comparison of the results. Since under-reporting does not appear to be problematic in both surveys (only 34 and 247 cases or less than 1% of the post-election respondents for the both ANES and CCES), we focus on the analysis of over-reporting rather than misreporting.

Following the literature on over-reporting (e.g., Ansolabehere and Hersh, 2012), we focus on the sample of validated non-voters only, which includes those respondents classified as non-voters in the 2016 Presidential Election by our validation procedure with fastLink and clerical review (1,390 and 21,835 respondents for the ANES and CCES, respectively). Figure 4 presents the

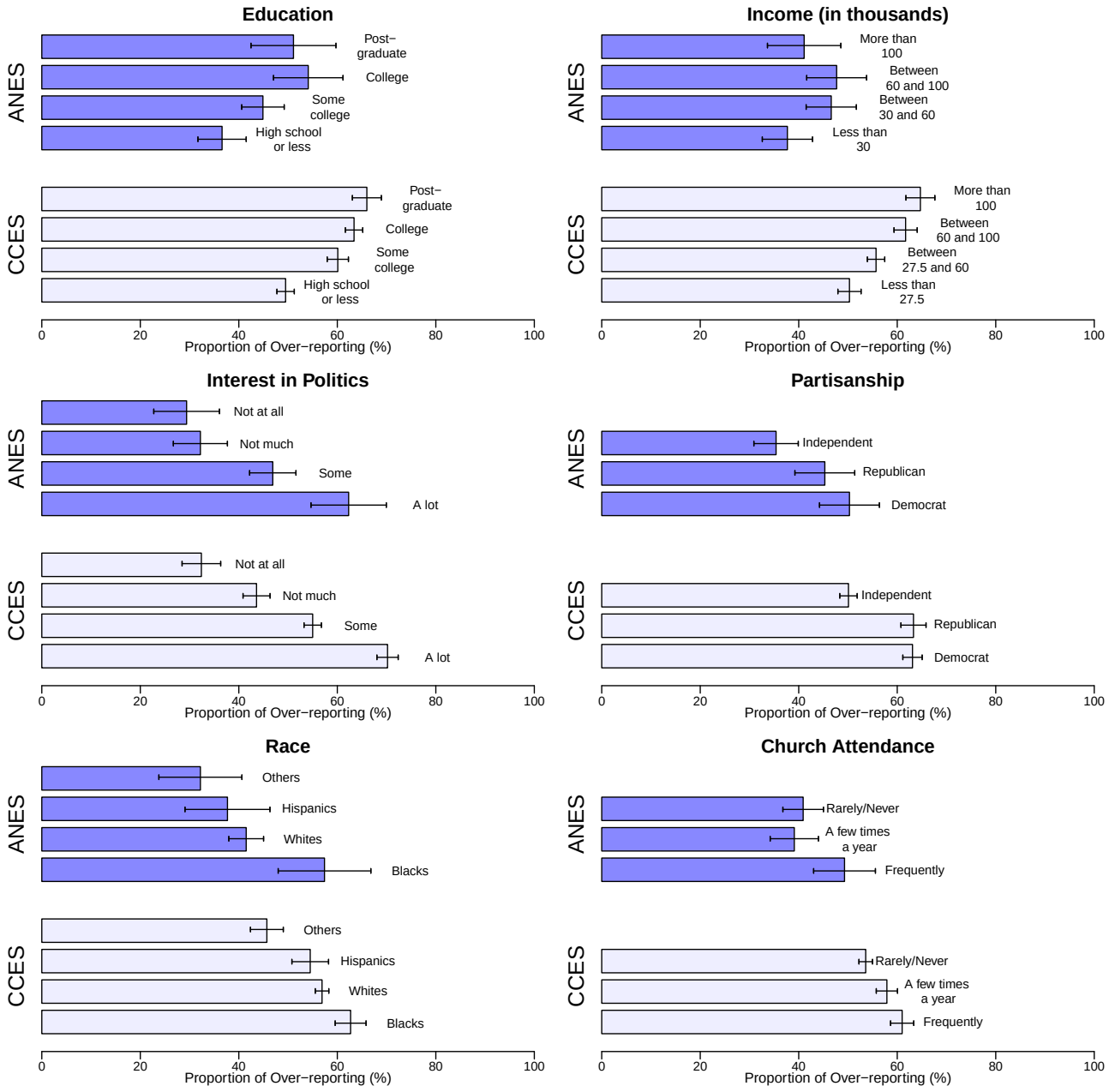


Figure 4: Estimated Proportion of Over-reporting across Different Covariates in the Sample of Validated Non-Voters. The results are based on the weighted logistic regression separately fitted to the CCES (light blue) and ANES (dark blue) samples of validated non-voters. Each plot presents the estimated proportion of over-reporting averaging over the entire sample of validated non-voters while fixing the other covariates at their observed values. Nonresponse is treated as a separate category for each covariate.

		Validation comparison			
		Common matches	Proprietary only	fastLink only	Overall
Validated turnout	fastLink	70.34 (0.35)	8.63 (0.21)	23.16 (0.43)	54.11 (0.31)
	Proprietary	68.48 (0.35)	10.14 (0.23)	0.00	52.85 (0.32)
Number of Observations		34,344	8,773	6,678	64,600

Table 7: Comparison of the Turnout Validation by fastLink and the Proprietary Validation Procedure Using the CCES Pre-election Sample. The table compares the validated turnout rates for three different groups of respondents: those declared as matches by both fastLink and the proprietary validation procedure ("Common matches"), those identified by the proprietary procedure only, and those matched by fastLink only.

estimated proportions of over-reports among the validated non-voters across the different values of some covariates, whose coefficients are estimated to be statistically significantly different from zero. These estimates are obtained by averaging over all respondents in the sample of validated non-voters (using the sampling weights) while fixing the other covariates to their observed values. Here, we graphically summarize the results, while the estimated coefficients and their standard errors are given in Table 10 of Appendix A.4.

For both the ANES and CCES, we find similar patterns: educated respondents tend to lie more than the uneducated, partisans are more likely to misreport than independents, and those who said they were interested in politics over-report more than those with little interest. Although the overall pattern is similar between the two surveys, there are some small differences. For example, for the CCES, there is a monotonic relationship between income and over-reporting: respondents with high income tend to over-report more than poor respondents. However, for the ANES, the relationship is not monotonic. In addition, for the ANES, we find a substantial difference in the propensity to over-report turnout between blacks and the other voters whereas the magnitude of this difference is much smaller for the CCES. These results are in line with the findings of other validation studies (see Ansolabehere and Hersh, 2012, and reference therein). As expected, we find that the CCES respondents tend to over-report turnout more than the ANES respondents.

4.4 Comparison with a Proprietary Algorithm

We compare the results of our algorithm with those of a proprietary algorithm. The CCES data set includes a validated turnout variable, which is produced by YouGov who used the voter file from another commercial firm, called Catalist. We use the updated validation results from the CCES as its initial version contained errors for many North Eastern states. We note that the results presented in this section should be interpreted with caution because the two algorithms are applied to two different national voter files. Although these voter files are based on the same data source, the differences in the results shown below may reflect those of voter files as well as those of the algorithms. Table 7 presents the validated turnout rates according to fastLink and the proprietary method for three different groups of respondents using the pre-election sample: those declared as matches by both fastLink and the proprietary method ("Common matches"), those identified by the proprietary method only, and those matched by fastLink only.

As expected, we find that the validated turnout rates are the highest among those who are matched to registered voters in the voter file by both fastLink and the proprietary method. Interestingly, while the matches identified only by the proprietary method have similarly low validated turnout rates according to both fastLink and the proprietary method, the validated turnout rate (according to fastLink) is much higher among the respondents whom only fastLink is able to match with registered voters. We note that the validated turnout rate for the respondents whom only proprietary method identified as matches is not zero according to fastLink because unlike the proprietary method fastLink allows unmatched respondents to have a positive posterior probability of match. Finally, the proprietary method underestimates the actual turnout rate by about 6 percentage points whereas the bias of fastLink is between 4 and 5 percentage points.

Figure 5 compares the accuracy of validated turnout rates at the state level using the pre-election sample. In each plot, the horizontal axis represents the actual turnout rate based on the voter file, whereas the vertical axis represents the validated turnoutq rate either based on the proprietary method (left plot) or fastLink (right plot). In particular, the bias, root mean squared error (RMSE), and correlation for fastLink are remarkably similar to those for the proprietary method. In sum, we find that at the aggregate level, fastLink performs at least as well as a state-of-art proprietary method.

Finally, we examine the differences in the individual level matching results of merging algo-

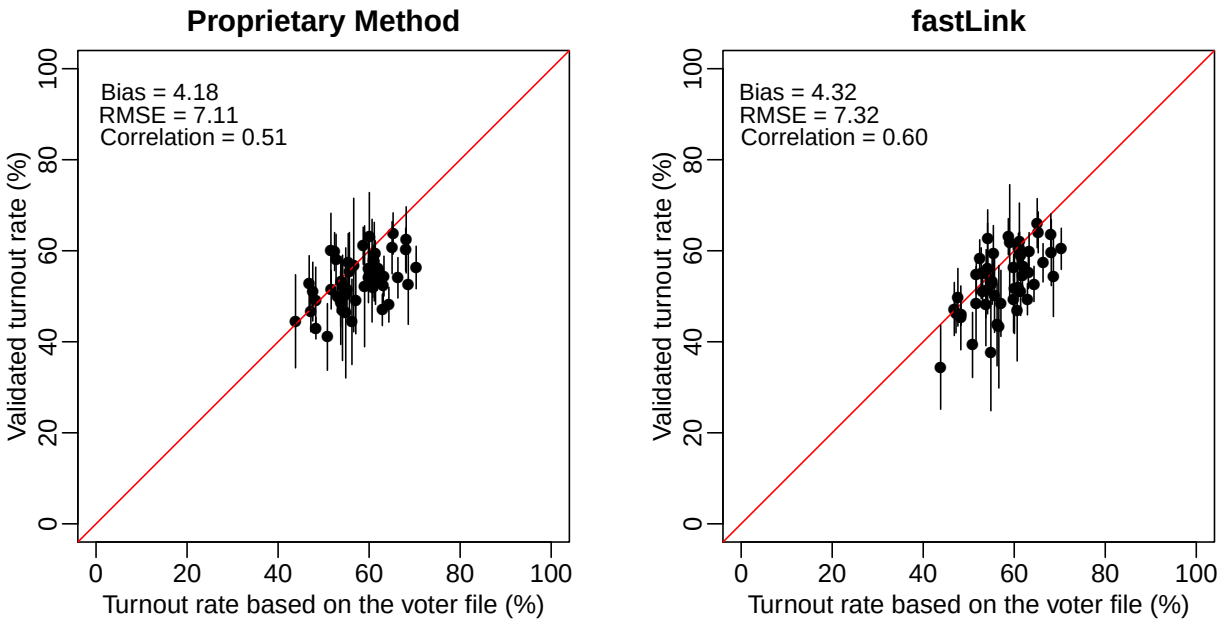


Figure 5: Comparison of the Validated Turnout Rates against the Actual Turnout Rates at the State-level. We evaluate the performance of the proprietary method (left plot) and fastLink (right plot) by plotting the resulting state-level validated turnout rate (vertical axis) against the actual turnout rate based on the voter file (horizontal axis).

rithms by conducting a regression analysis using the post-election sample of the CCES. In our analysis, the outcome variable takes four categorical values: matched by both fastLink and the proprietary algorithm, matched by neither algorithm, matched only by fastLink, and matched only by the proprietary algorithm. Using this outcome variable, we fit a weighted multinomial logistic regression model with survey weights, and include the same set of covariates used to predict over-reporting. The estimated coefficients and their standard errors are given in Table 11 of Appendix A.6, which in addition contains a complete description how the estimation was conducted.

Figure 6 presents the predicted probabilities for four possible matching statuses based on the fitted weighted multinomial logit model. We compute the predicted probability by setting a covariate to a specific value and averaging over the entire post-election sample (52,899 observations) while fixing the other covariates at their observed values. The figure presents the results for a subset of covariates. Overall, the proprietary algorithm finds a few more matches than fastLink. We find that hispanic voters are likely to be unmatched when compared to black and white voters. Not surprisingly, the probability of being matched by both algorithms is greater for the voters who are interested in politics and have higher income. Interestingly, the proprietary algorithm ends up matching more lower income individualss.

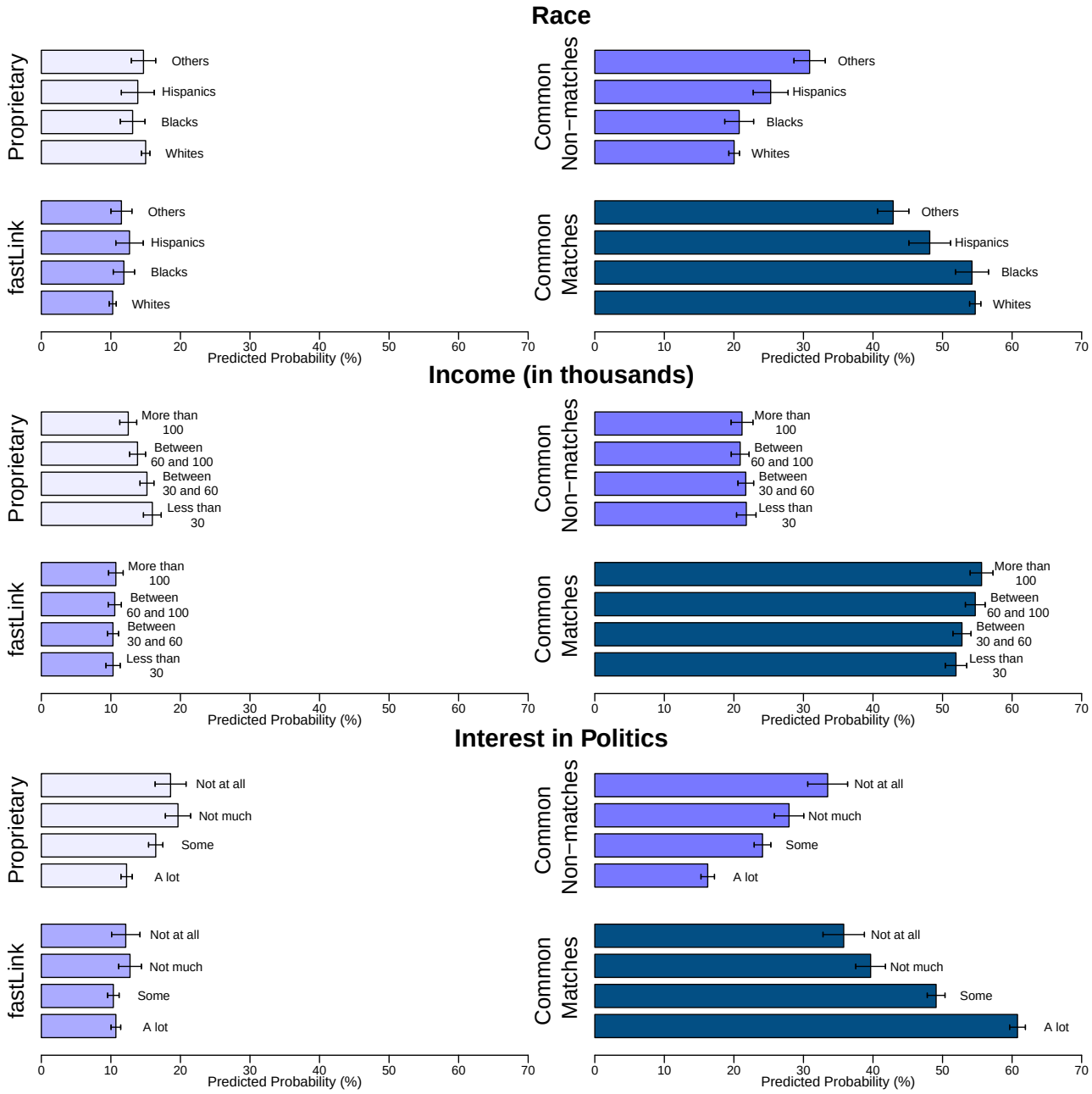


Figure 6: Predicted Probabilities for the Matching Status of the Different Vote Validation Exercises across Covariates. The results are based on the weighted multinomial logistic regression, where the outcome variable takes four values, indicating different matching status for each CCES respondent: matched by both fastLink and the proprietary algorithm (dark blue), matched by neither algorithm (medium blue), matched only by fastLink (light blue), and matched only by proprietary algorithm (white). Each plot presents the estimated predicted probabilities averaging over the entire post-election sample while fixing the other covariates at their observed values. Nonresponse is treated as a separate category for each covariate.

5 Concluding Remarks

Over the last decade, the availability of large-scale electronic administrative records enabled researchers to study important questions by creatively merging them with other data sets (see e.g., Jutte, Roos and Browne, 2011; Ansolabehere and Hersh, 2012; Einav and Levin, 2014). A major methodological challenge of these studies, however, is that there often exists no unique identifier that can be used to unambiguously merge data sets. In these situations, probabilistic record linkage methods that have been developed in the statistics literature over the last several decades can serve as a useful methodological tool. This paper presents a case study that applies the canonical record linkage method of Fellegi and Sunter (1969) to merge two prominent national election survey data sets with the nationwide voter file of more than 180 million records. We show that the recent computational improvements makes it possible to conduct this large-scale data merge. Our analysis demonstrates that the probabilistic record linkage method can successfully validate turnout and shed light on the debate regarding the potential causes of bias in self-reported turnout. The probabilistic method is especially effective dealing with missing and invalid entries as shown in the case of the CCES validation. We believe that a similar application of probabilistic record linkage methods in other domains can also be fruitful, leading to new scientific discoveries.

References

American Association for Public Opinion Research. 2017. An Evalution of 2016 Election Polls in the United States. Technical Report. Ad Hoc Committee on 2016 Election Polling.

American National Election Studies. 2017. User Guide and Codebook for the ANES 2016 Time Series Study. Technical Report. University of Michigan and Stanford University Ann Arbor, MI and Palo Alto, CA: .

ANES. 2016. "User's Guide and Codebook for the ANES 2012 Time Series Voter Validation Supplemental Data." Ann Arbor, MI and Palo Alto, CA: the University of Michigan and Stanford University.

Ansolabehere, Stephen, Brian Schaffner and Sam Luks. 2017. Guide to the 2016 Cooperative Congressional Election Survey. Technical Report. Harvard University. Data Release No. 2.

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

Berent, Matthew K., Jon A. Krosnick and Arthur Lupia. 2011. The Quality of Government Records and "Overestimation" of Registration and Turnout in Surveys: Lessons from the 2008 ANES Panel Studys Registration and Turnout Validation Exercises. Technical Report No. nes012554. American National Election Studies Ann Arbor, Michigan and Palo Alto, California: .

Berent, Matthew K., Jon A. Krosnick and Arthur Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys." *Public Opinion Quarterly* 80:597–621.

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

Burden, Barry C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8:389–398.

DeBell, Matthew, Michelle Amsbary, Vanessa Meldener, Shelly Brock and Natalya Maisel. 2016. Methodology Report for the ANES 2016 Time Series Study. Technical Report. ANES, Stanford University and the University of Michigan.

Einav, Liran and Jonathan Levin. 2014. "Economics in the age of big data." *Science* 346.

Enamorado, Ted, Benjamin Fifield and Kosuke Imai. 2017a. "fastLink: Fast Probabilistic Record Linkage." available at the Comprehensive R Archive Network (CRAN). `https://CRAN.R-project.org/package=fastLink`.

Enamorado, Ted, Benjamin Fifield and Kosuke Imai. 2017b. Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records. Technical Report. Department of Politics, Princeton University.

Enns, Peter K., Julius Lagodny and Jonathon P. Schuldt. 2017. "Understanding the 2016 US Presidential Polls: The Importance of Hidden Trump Supporters." *Statistics, Politics, and Policy* 8:41–63.

Fellegi, Ivan P. and Alan B. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64:1183–1210.

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

Jackman, Simon and Bradley Spahn. 2017. *Why Does the American National Election Study Overestimate Voter Turnout?* Department of Political Science, Stanford University.

Jutte, Douglas P., Leslie L. Roos and Marni D. Browne. 2011. "Administrative Record Linkage as a Tool for Public Health Research." *Annual Review of Public Health* 32:91–108.

Lahiri, P. and Michael D. Larsen. 2005. "Regression Analysis with Linked Data." *Journal of the American Statistical Association* 100:222–230.

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

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

Winkler, William E. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." Proceedings of the Section on Survey Research Methods. American Statistical Association.

Winkler, William E. 1995. *Business Survey Methods*. New York: J. Wiley Chapter Matching and Record Linkage, pp. 355–84.

Winkler, William E. 2006. Overview of record linkage and current research directions. Technical Report. United States Bureau of the Census.

A Supplementary Appendix

A.1 Turnout and Registration Questions in the 2016 ANES and CCES

In this appendix, we present the question wording and coding rules used for self-reported registration and turnout.

A.1.1 Self-reported Registration

For the 2016 ANES, a respondent was asked the following question regarding registration in both pre-election (`V161011`) and post-election (`V162022`) surveys.

```
Are you
1. registered to vote at this address?
2. registered at a different address?
3. not currently registered
```

We code answers 1 and 2 as *registered* and answer 3 as *not registered* to vote in the 2016 General election. Only those who gave answer 3 in the pre-election survey were asked the registration question again in the post-election survey. Thus, in our analysis, we treat the respondents who said in the pre-election survey they had registered as registered in the post-election survey as well.

Similarly, the CCES asks respondents about their registration status in both pre-election and post-election surveys. The variable that summarizes the registration status for all the respondents in the CCES is `votereg`. The question reads, `Are you registered to vote?`, and the possible answers are `yes, no, or don't know`.

A.1.2 Self-reported Turnout

In the ANES pre-election survey, respondents were asked, `Did you vote for President in 2016?`, with yes and no as the possible answers (`V161026`). This question is designed to capture early and absentee voting, and the respondents who answered yes to this question are not asked again the turnout question in the post-election survey. The self-reported turnout question for the post-election survey (`V162031`) reads as follows:

```
Which of the following statements best describes you?
1. I did not vote (in the election this November).
```

2. I thought about voting this time, but didn't.

3. I usually vote, but didn't this time.

4. I am sure I voted.

The information in both V161026 and V162031 is summarized by the ANES as V161026x, which we analyze. This variable is constructed as follows: respondents who gave an answer other than option 4 for V162031, or declared that they did not vote early are coded as non-voters. Those who gave answer 4 in V162031 or declared that they had voted early in V161026 are coded as voters.

The equivalent post-election question in the CCES (CC16_401) uses a similar wording. The question reads:

Which of the following statements best describes you?

1. I did not vote in the election this November.

2. I thought about voting this time but didn't.

3. I usually vote, but didn't this time.

4. I attempted to vote but did not or could not.

5. I definitely voted in the General Election.

Question CC16_401 is asked to every respondent of the post-election survey, regardless of whether they have declared to have voted early or cast absentee ballots in the pre-election survey (1,521 respondents). The turnout question in the pre-election survey (CC16_364) reads, Do you intend to vote in the 2016 general election? For our analysis, we use CC16_401, which represents respondents' most recent recollection of turnout decision. We code as non-voters the respondents who chose answers 1 through 4, and as voters those who chose answer 5.

A.2 Sampling Weights

When incorporating the sampling design of the 2016 ANES in our analyses, the following variables are used as sampling weights:

- Overall sample:
  - Primary sampling unit (PSU): V160202
  - Stratum: V160201

  - Weights: `V160101` (pre-election), `V160102` (post-election)

- Face-to-face sample:
  - Primary sampling unit (PSU): `V160202f`
  - Stratum: `V160201f`
  - Weights: `V160101f` (pre-election), `V160102f` (post-election)

- Internet sample:
  - Primary sampling unit (PSU): `V160202w`
  - Stratum: `V160201w`
  - Weights: `V160101w` (pre-election), `V160102w` (post-election)

For the 2016 CCES, we use the following variables as sampling weights in our analyses:

- Weights: `commonweight` (pre-election), `commonweight_post` (post-election)

The CCES conducts their own turnout validation, and recalibrate the weights to match the CPS estimates. However, we do not use `commonweight_vv` (pre-election) and `commonweight_post_vv` (post-election) as sampling weights because they are based on the CCES turnout validation. Instead, we use `commonweight` (pre-election) and `commonweight_post` (post-election), which were constructed before the turnout validation was performed and hence are appropriate for a fair comparison of the CCES and fastLink turnout validation.

A.3 Empirical Results without Sampling Weights

		Pre-election		Post-election		Actual turnout	
		fastLink	clerical review	fastLink	clerical review	Voter file	Election project
ANES	Overall	65.41 (0.72)	60.17 (0.75)	66.75 (0.77)	61.88 (0.80)	57.55	58.83
	Internet	65.47 (0.84)	60.26 (0.88)	65.67 (0.91)	60.48 (0.95)	57.55	58.83
	Face-to-face	65.26 (1.36)	59.95 (1.43)	66.51 (1.43)	61.95 (1.49)	57.58	58.86
CCES		58.24 (0.19)	52.82 (0.20)	63.09 (0.21)	58.17 (0.21)	57.55	58.83

Table 8: Validated Turnout Rates among the Survey Respondents from the ANES and CCES. The validated turnout rates obtained from the probabilistic model alone ("fastLink") and the model plus clerical review are compared to the actual turnout rate for the corresponding target population based on the voter file and the data from the United States election project. The standard errors are given in parentheses. These results do not make use of sampling weights.

			Registered		Post-election
		Not registered	Did not Vote	Vote	attrition
ANES	fastLink	8.16 (1.36)	14.63 (1.58)	81.87 (0.71)	57.55 (1.95)
	Clerical review	0.58 (0.41)	7.21 (1.23)	77.71 (0.78)	50.16 (2.01)
	Number of respondents	342	444	2862	622
CCES	fastLink	12.48 (0.44)	9.35 (0.47)	72.42 (0.21)	36.31 (0.43)
	Clerical review	4.18 (0.29)	3.80 (0.34)	67.91 (0.22)	28.63 (0.42)
	Number of respondents	4684	3237	44796	11701

Table 9: Validated Turnout Rates among Survey Respondents with Different Responses to the Turnout Questions in the ANES and CCES. "Post-election attrition" refers to the group of survey respondents who did not answer the turnout questions due to attrition. Standard errors are in parentheses. These results do not make use of sampling weights.

A.4 Regression Models for Over-reporting

In this appendix, for each survey, we present two sets of estimated coefficients for the weighted logistic regression using survey weights. The models are fitted to the sample of validated non-voters alone. Specifications (1) and (3) in Table 10 present the estimates obtained by fitting the weighted logistic regression models to the sample of validated non-voters, in which non-response is coded as a separate category for each variable to account for missing values. These specifications do not drop any observation and form the basis of the graphical summaries presented in Section 4.3. The second set of estimates, shown as Specifications (2) and (4) in Table 10, perform a similar analysis but use listwise deletion to deal with missing values.

A.5 Description of Variables Used to Predict Over-reporting

- **Age** measured in years since the date of birth.
  - ANES: `V161267` respondent's age.
  - CCES: `birthyr` year of birth.
- **Marital status** collapsed into three distinct categories: Married, Widowed/Divorced, and Never married.
  - ANES: `V161268` marital status. Married = { Married: spouse present, Married: spouse absent}, Widowed/Divorced = { Widowed, Divorced }, Never married = { Never married }.
  - CCES: `marstat` marital status. Married = { Married, Domestic partnership }, Widowed/Divorced = { Widowed, Separated, Divorced }, Never married = { Single }.
- **Education** collapsed into four categories: High school or less, Some college, College, and Post-graduate.
  - ANES: `V161270` Highest level of Education. High school or less = { values less than 10 }, Some college = { values between 10 and 12}, College = { 13 }, Post-graduate = { values between 14 and 16 }.

	ANES				CCES			
	(1)		(2)		(3)		(4)	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Age:								
Years of age	0.010	0.028	−0.035	0.032	−0.044	0.012	−0.041	0.014
Years of age squared	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000
No response	0.767	0.759						
Marital Status:								
Widowed/Divorced	−0.431	0.287	−0.466	0.344	−0.216	0.085	−0.209	0.095
Never married	−0.061	0.205	−0.049	0.256	−0.071	0.078	0.025	0.089
No response	−0.095	0.205			2.194	0.508		
Education:								
Some College	0.414	0.181	0.269	0.264	0.567	0.078	0.539	0.090
College	0.860	0.230	0.625	0.289	0.754	0.072	0.713	0.084
Post-graduate	0.716	0.244	0.751	0.302	0.905	0.100	0.815	0.115
No response	−0.287	0.793						
Gender:								
Male	−0.052	0.131	0.009	0.196	0.278	0.062	0.344	0.070
No response	1.134	0.788						
Race:								
Black	0.788	0.275	0.817	0.421	0.335	0.105	0.293	0.116
Hispanic	−0.193	0.246	−0.419	0.316	−0.133	0.114	−0.130	0.130
Other	−0.491	0.263	−0.721	0.319	−0.608	0.100	−0.639	0.115
No response	0.572	1.053						
Income (in thousands):								
Between 27.5 and 60	0.455	0.211	0.854	0.299	0.285	0.080	0.343	0.088
Between 60 and 100	0.507	0.211	0.620	0.302	0.620	0.092	0.676	0.100
More than 100	0.173	0.265	0.441	0.358	0.791	0.112	0.854	0.124
No response					0.356	0.111		
Partisanship:								
Republican	−0.242	0.235	−0.202	0.274	0.012	0.094	0.047	0.103
Independent	−0.736	0.188	−0.459	0.224	−0.682	0.075	−0.674	0.082
No response	−1.039	0.447			−1.351	0.140		
Interest in Politics:								
Some	−0.723	0.229	−0.749	0.284	−0.795	0.075	−0.762	0.084
Not very	−1.435	0.240	−1.610	0.339	−1.345	0.091	−1.379	0.103
Not at all	−1.588	0.264	−1.339	0.394	−1.908	0.122	−2.116	0.144
No response	−0.446	0.841			−1.478	0.234		
Church Attendance:								
A few times a year	−0.514	0.193	−0.417	0.254	−0.175	0.092	−0.212	0.105
Rarely/Never	−0.422	0.193	−0.413	0.236	−0.411	0.079	−0.419	0.091
No response	−0.355	0.775			−0.469	0.282		
Ideology:								
Liberal	−0.217	0.369	−0.354	0.415	−0.088	0.136	−0.134	0.147
Moderate	−0.509	0.257	−0.551	0.284	−0.077	0.129	−0.114	0.140
Conservative	−0.251	0.282	−0.431	0.320	−0.107	0.139	−0.215	0.150
Very conservative	−0.286	0.304	−0.241	0.324	−0.322	0.170	−0.331	0.188
No response	−0.614	0.256			−0.933	0.167		
Intercept	0.692	0.751	1.475	0.848	1.912	0.318	1.899	0.359
Number of observations:	1,390		772		21,835		16,554	

Table 10: Estimated Coefficients for the Weighted Logistic Regression of Over-reporting. The estimates presented here, and their corresponding standard errors, are obtained from a logistic regression adjusting by the sampling design of the ANES and the CCES. Specifications (1) and (3) refer to the results obtained when coding nonresponse as a separate category. Specifications (2) and (4) refer to the results obtained when listwise deletion is applied for missing values.

  - CCES: `educ` What is the highest level of education you have completed? High school or less = { No High school }, Some college = { Some college }, College = { 2-year, 4-year }, Post-graduate = { Post-grad }.

- **Gender.** equals to 1 for males and 0 for females.
  - ANES: `V161002` gender.
  - CCES: `gender` Are you male or female?

- **Race.** which is collapsed into four categories: White, Black, Hispanic, and Other.
  - ANES: `V161310x` self-identified race. White = { White, non-Hispanic }, Black = { Black, non-Hispanic }, Hispanic = { Hispanic }, Other = { Asian, native Hawaiian or other Pacific Iscalnder, non-Hispanic, Native American or Alaska native, non-Hispanic, Other non-Hispanic including multiple races }.
  - CCES: `race` What racial or ethnic group best describes you? White = { White }, Black = { Black }, Hispanic = { Hispanic }, Other = { Asian, Native American, Middle Eastern, Mixed, Other }.

- **Income.** collapsed into four categories: Less than 30 thousand, Between 30 and 60 thousand, Between 60 and 100 thousand, more than 100 thousand.
  - ANES: `V161361x` Income summary.
  - CCES: `faminc` Thinking back over the last year, what was your family's annual income?

- **Partisanship.** party affiliation, collapsed into three categories: Democrat, Republican, and Independent.
  - ANES: `V161155` Does R think of self as Dem, Rep, Ind or what? Democrat = { Democrat }, Republican = { Republican }, Independent = { Independent, Other }.
  - CCES: `pid3` Generally speaking, do you think of yourself as a ...? Democrat = { Democrat }, Republican = { Republican }, Independent = { Independent, Other }.

- **Interest in Politics.** collapsed into the following four categories: A lot, Some, A little, None.

- ANES: `V162256` respondent's interest in politics. A lot = { Very interested }, Some = { Somewhat interested }, A little = { Not very interested }, None = {Not at all interested }.
- CCES: `newsint` Interest in politics. Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say you follow what's going on in government and public affairs...? A lot = { Most of the time }, Some = {Some of the time }, A little = { Only now and then }, None = { Hardly at all }.

- **Religiosity.** which is proxied here by church attendance and collapsed as follows: Frequently, A few times a year, Rarely/Never.
  - ANES: `V161245` Attend religious services how often. Frequently = { Every week, Almost every week }, A few times a year = { Once or twice a month, A few times a year }, Rarely/Never = { Never in V161245, and No in V161244}. Where `V161244` asks Do you ever attend church or religious services?
  - CCES: `pew_churatd` Aside from weddings and funerals, how often do you attend religious services? Frequently = { More than once a week, Once a week }, A few times a year = { Once or twice a month, A few times a year }, Rarely/Never = { Seldom, Never}.

- **Ideology.** collapsed into five categories: Very liberal, Liberal, Moderate, Conservative, Very conservative.
  - ANES: `V161126` 7pt scale Liberal conservative self-placement.
  - CCES: `ideo5` In general, how would you describe your own political viewpoint?

A.6 Additional Results: Comparison with a Proprietary Algorithm

To define our dependent variable (match status), note first that according to fastLink each observation has a probability equal to ζ_i of being a match and $1-\zeta_i$ of being a non-match. The proprietary method on the other hand, either declares observations matches or non-matches. Thus, construct the four mutually exclusive categories describing match status, we use the following approach:

1. Create a duplicate for each observation in the CCES.

2. Assign a fictitious fastLink of non-match status to each duplicate, while the opposite status (match) will be assigned to the original observations.

3. Use the observed the proprietary method along with the fictitious fastLink matching status to classify observations into one of the following groups:

4. Fit a weighted multinomial logistic regression on the new data, fixing the weights to equal the product between the observed ζ_i $(1 - \zeta_i)$ and the sample weights for each observation if the fictitious fastLink status is equal to match (non-match).

Such an approach has the advantage that for each observation in the CCES it exploits all the information contained in ζ_i.

Note that the standard errors for the predicted probabilities presented in Figure 6 need to adjust for the inclusion of ζ_i and $1-\zeta_i$ in the estimation stage. To approximate the standard error of each predicted probability we make use of non-parametric bootstrap. We draw 1,000 samples with replacement, fit the weighted multinomial logistic regression delineated above, and produce estimates for the predicted probabilities of interest.

	Common Non-Matches		fastLink only		Common Matches	
	est.	s.e.	est.	s.e.	est.	s.e.
Age:						
Years of age	−0.035	0.001	−0.019	0.002	−0.010	0.001
Years of age squared	0.000	0.000	0.000	0.000	0.000	0.000
Marital Status:						
Widowed/Divorced	0.021	0.006	0.076	0.004	−0.050	0.011
Never married	−0.030	0.011	0.205	0.008	0.218	0.014
No response	1.293	0.000	−0.274	0.000	0.334	0.000
Education:						
Some College	−0.183	0.011	0.192	0.007	0.253	0.015
College	−0.002	0.011	0.260	0.008	0.287	0.014
Post-graduate	0.218	0.005	0.413	0.003	0.277	0.009
Gender:						
Male	0.254	0.014	0.223	0.012	0.148	0.014
Race:						
Black	0.173	0.006	0.281	0.003	0.124	0.008
Hispanic	0.335	0.001	0.288	0.001	−0.080	0.001
Other	0.497	0.002	0.130	0.001	−0.283	0.003
Income (in thousands):						
Between 30 and 60	0.045	0.012	0.049	0.008	0.070	0.013
Between 60 and 100	0.094	0.009	0.169	0.006	0.208	0.014
More than 100	0.007	0.286	0.004	0.329	0.012	
No response	0.200	0.004	0.437	0.001	0.133	0.005
Partisanship:						
Republican	−0.258	0.008	−0.313	0.006	−0.162	0.010
Independent	−0.013	0.010	−0.086	0.008	−0.266	0.012
No response	0.187	0.006	−0.070	0.003	−0.781	0.003
Interest in Politics:						
Some	0.131	0.010	−0.324	0.004	−0.534	0.013
Not very	0.119	0.004	−0.289	0.002	−0.943	0.005
Not at all	0.372	0.004	−0.278	0.002	−1.000	0.002
No response	0.269	0.001	−0.339	0.000	−1.096	0.000
Church Attendance:						
A few times a year	−0.164	0.009	−0.189	0.008	−0.127	0.008
Rarely/Never	−0.229	0.011	−0.190	0.011	−0.102	0.010
No response	0.174	0.001	−0.437	0.000	−0.156	0.000
Ideology:						
Liberal	0.210	0.007	0.092	0.003	0.041	0.011
Moderate	0.263	0.011	0.125	0.009	−0.047	0.011
Conservative	0.370	0.009	0.190	0.008	−0.012	0.010
Very conservative	0.207	0.001	−0.013	0.001	0.005	0.002
No response	0.331	0.007	0.164	0.003	−0.276	0.004
Intercept	1.169	0.002	−0.007	0.001	1.333	0.002

Table 11: Estimated Coefficients for the Weighted Multinomial Logistic Regression of Validation Type. The estimates and their corresponding standard errors are obtained from a multinomial logistic regression adjusting by the sampling design of the CCES. Base category is Proprietary Method only. All the specifications refer to the results obtained when coding nonresponse as a separate category.