Supplementary Appendix for "Validating Self-reported Turnout by Linking Public Opinion Surveys with Administrative Records"

Ted Enamorado* Kosuke Imai†

A1 Additional Information about the 2016 ANES and CCES

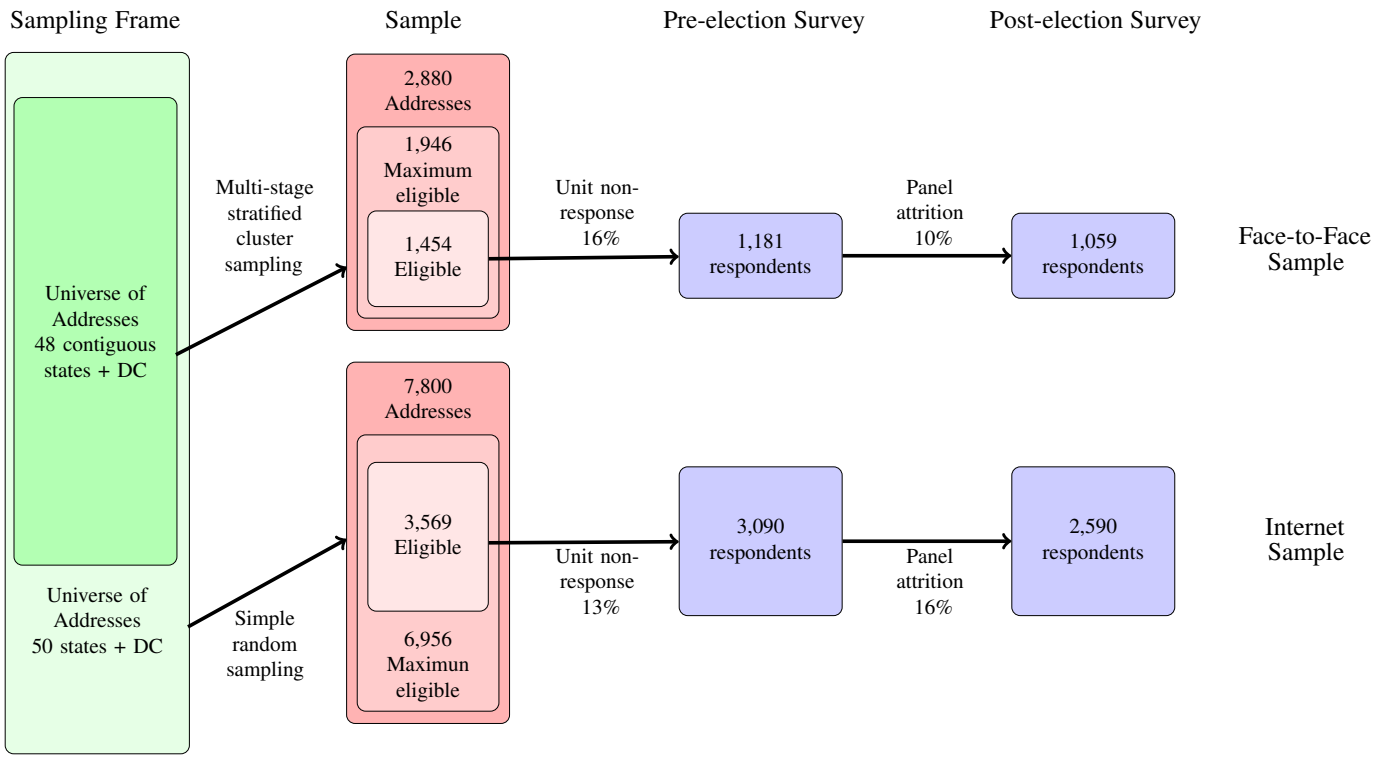
A1.1 The Sampling Designs of the ANES and CCES

Figures A1a and A1b 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 A1a, 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.

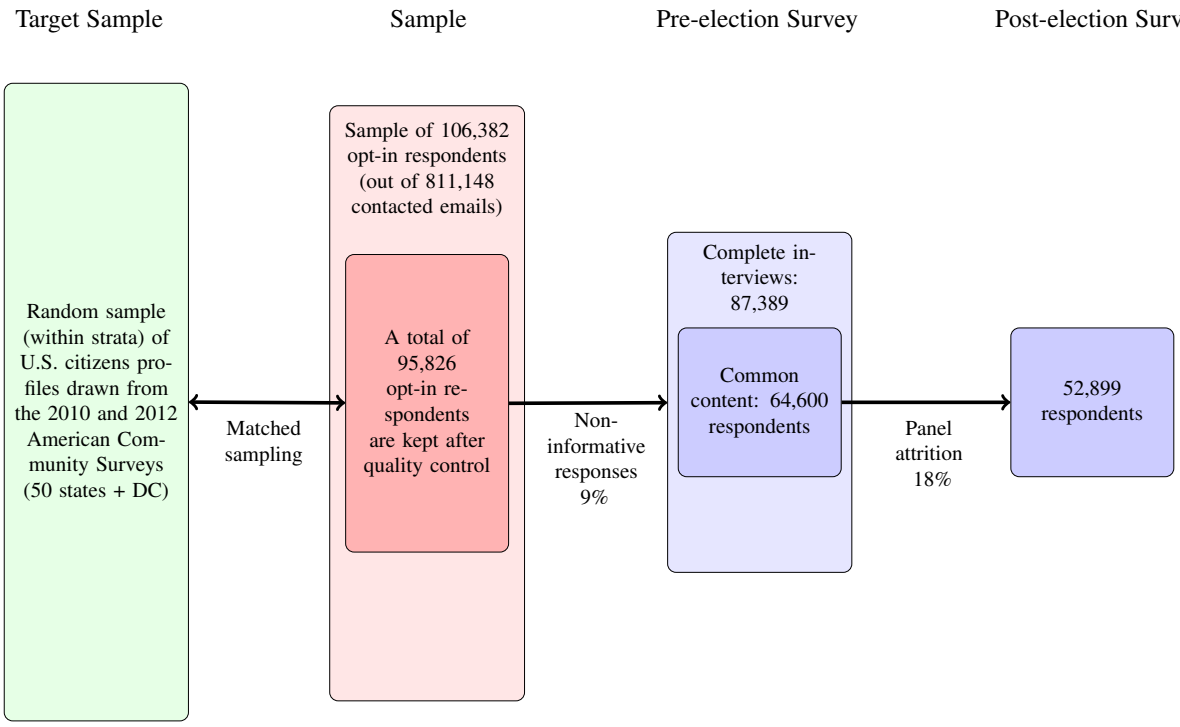
Finally, a trained interviewer was sent to each selected household and administered the survey to a randomly selected adult citizen. Thus, the target population is 222.6 million adult US citizens age 18 or older who reside in contiguous 48 states and D.C. There was a unit non-response rate of 16% in 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

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

†Professor of Government and of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu



(a) The Sampling Design of the 2016 ANES



(b) The Sampling Design of the 2016 CCES

Figure A1: 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.

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.[1] The bottom panel of Figure A1a 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 *et al.*, 2017, for details). As summarized in Figure A1b, the CCES first constructed a "target sample" from the respondents who had participated in the 2010 and 2012 American Community surveys and other 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 political 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

[1]This sample excludes 'drop points'. As defined in DeBell *et al.* (2016) a 'drop point' is an address associated with multiple dwellings and where the same mail box is used by more than one those dwellings.

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.

A1.2 Preprocessing of Names and Addresses

To parse the de-anonymized names of the respondents of the ANES and CCES into first, middle, and last name, we use the following steps:

1. Parse names into first, middle, and last name. We used the R package `humaniformat`.
2. Remove names suffixes. Due to its rare occurrence, suffixes were not considered for further parsing.
3. Classify as missing values cases with no information. In the case of the CCES, there were 1,748 records for which no usable information was provided to recover names. In the case of the ANES, only 82 such cases were found.
4. Double-check that the names were correctly parsed. We use the `python` module `probablepeople` was used. No discrepancies were found.

To parse the de-anonymized address for the CCES respondents, we use the following steps:

1. Standardize the information for each address. As noted above, we used `preprocText()` function in fastLink to standardize each address according to the USPS Postal Address Information System
2. Parse addresses into house number, street name, and zip code. We used the `python` module `usaddress`.
3. Classify as missing values cases with no information. There were 7,465 records for which no usable information was provided to recover an address.

A1.3 Turnout and Registration Questions

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

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`.

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.

A1.4 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.

A1.5 Description of Variables Used to Predict Overreporting

- **Age** measured in years since the date of birth and collapsed into four categories: 18-34, 35-44, 45-54, 55+.
  - 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 }.
  - 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?

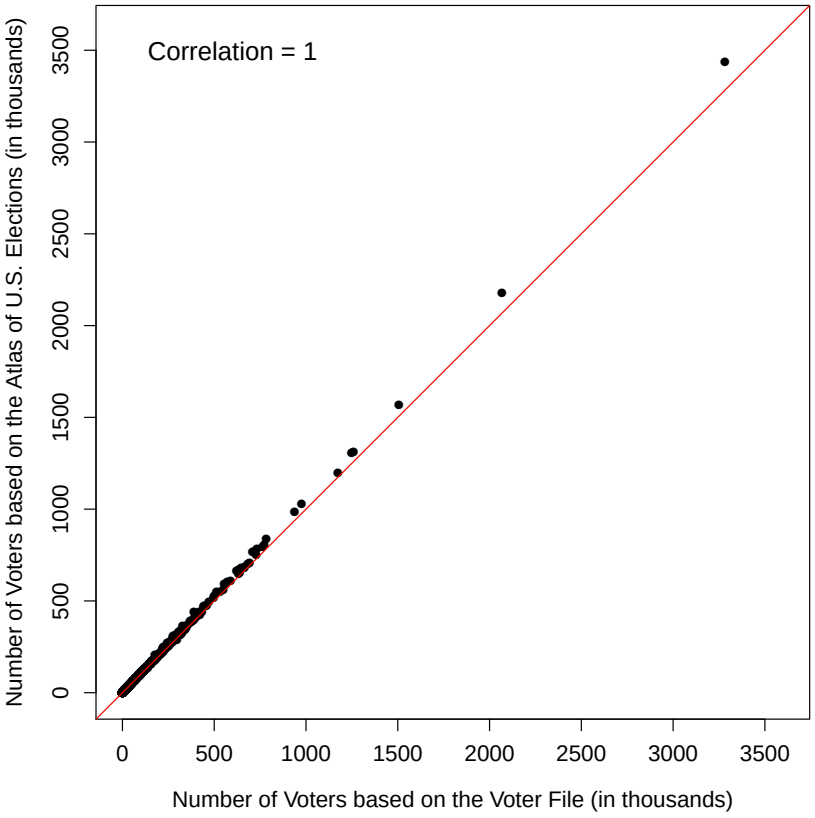


Figure A2: County-level Comparison between the Number of Voters based on the Voter File and the Atlas of U.S. Elections. While the number of voters from the Atlas of U.S. Elections tends to be slightly greater, there is a near perfect correlation.

A2 Additional Empirical Results

A2.1 County-level Comparison

Here, we further examine the accuracy of the L2 voter file. Specifically, Figure A2 compares the number of voters in the 2016 election at the county-level based on the L2 voter file with that from Dave Leip's Atlas of U.S. Presidential Elections (`https://uselectionatlas.org`), which has been collecting election results at the county-level for more than 20 years. In the specific case of the 2016 Presidential Election, the Atlas contains county-level information for every state with the exception of Alaska. We find that while the number of voters per county in the Atlas of U.S. Presidential Elections is slightly greater (4 percentage points on average) than the one obtained from the voter file, there is a near perfect correlation between the two measures. This suggests that the L2 voter file does not have a systematic bias.

		Pre-election (fastLink)		Post-election (fastLink)		Actual turnout	
		one-to-one match	one-to-many match	one-to-one match	one-to-many match	Voter file	Election project
ANES	Overall	63.59 (0.91)	63.81 (0.88)	64.96 (0.96)	65.06 (0.91)	57.55	58.83
	Internet	62.59 (1.06)	62.72 (1.03)	63.99 (1.15)	63.92 (1.10)	57.55	58.83
	Face-to-face	66.46 (1.76)	66.97 (1.63)	67.59 (1.69)	68.15 (1.53)	57.58	58.86

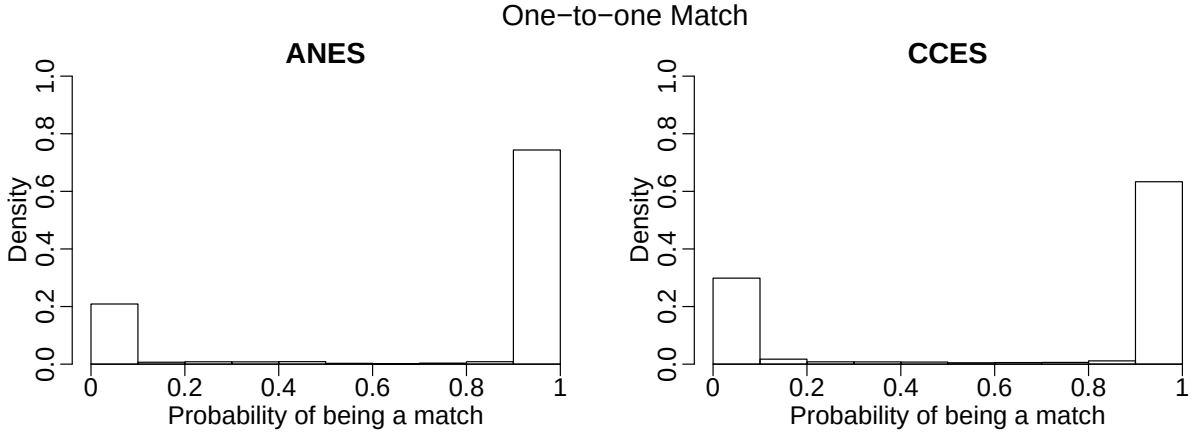
Table A1: Validated Turnout Rates among the Survey Respondents from the 2016 ANES: one-to-one vs one-to-many match. The validated turnout rates obtained from the probabilistic model alone ("fastLink"), under a one-to-one and one-to-many matching restriction, are compared to the actual turnout rate for 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.

A2.2 Estimated Turnout Rates for the ANES Using One-to-Many Matching Strategy

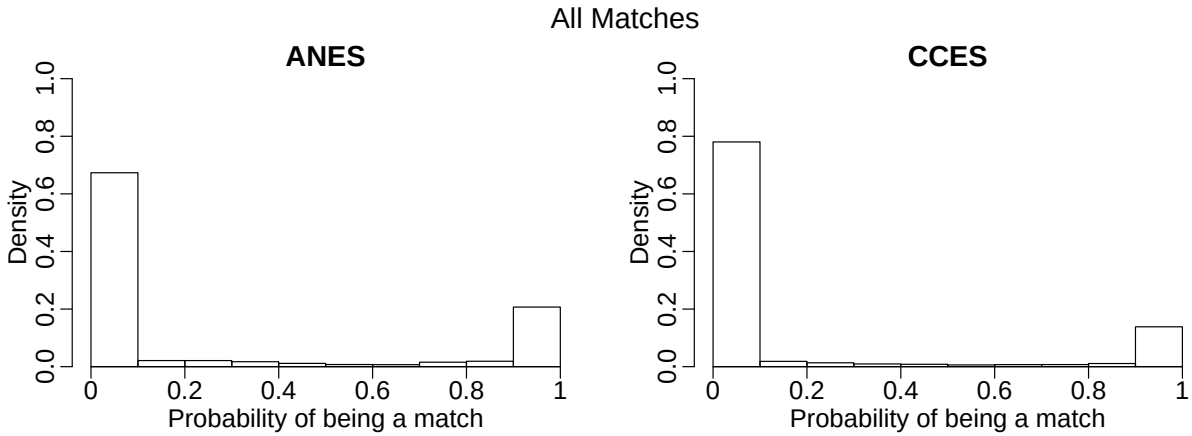
As described in Section 3.2, we use a one-to-one merge. Here, we examine the robustness of the results by applying the one-to-many matching strategy as described in Enamorado *et al.* (2019). Specifically, we compute the weighted average of turnout variable among all the matched records in the voter file, where the weights are proportional to the match probabilities. Formally, for each respondent i, we compute $\sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} T_j / \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij}$, where T_j represents a binary turnout variable for registered voter j in the voter file, ξ_{ij} is the estimated probability of respondent i being matched with voter j, and $N_{\mathcal{B}}$ is the number of records in the voter file. Table A1 compares the results based on the one-to-one and one-to-many matching strategies. The results are essentially identical regardless of interview mode.

A2.3 Distribution of the Estimated Match Probabilities

Figure A3 presents the distribution of the match probabilities for the ANES and CCES. Figure A3a shows that the probabilistic model after a one-to-one matching restriction separates the data quite well into matches and non-matches. Figure A3b shows that a similar pattern is found even when examining all matches, suggesting that the difference between one-to-one matching and one-to-many matching is minimal.



(a) One-to-one match



(b) One-to-many match

Figure A3: Distributions of the Estimated Match Probabilities for the ANES and CCES

A2.4 Relaxing the Conditional Independence Assumption

In this appendix, we follow the literature and use log-linear models to relax the conditional independence assumption. This approach can account for general patterns of dependence across variables (see e.g., Winkler, 1989, 1993; Thibaudeau, 1993; Larsen and Rubin, 2001, and references therein). Here, we show that the resulting matched and validated turnout rates are somewhat more conservative than those obtained under the conditional independence assumption.

Formally, the observed-data likelihood function of the Fellegi-Sunter model *without* the conditional

independence assumption is given by,

$$\mathcal{L}_{obs}(\lambda, \boldsymbol{\theta} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) \;=\; \prod_{i=1}^{N_{\mathcal{A}}} \prod_{j=1}^{N_{\mathcal{B}}} \left\{ \sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \boldsymbol{\pi}_m(i, j; \boldsymbol{\theta}_m) \right\}$$

where $\boldsymbol{\pi}_m(i,j) = \Pr(\gamma(i,j) \mid M_{ij} = m, \boldsymbol{\theta}_m)$ for $m \in \{0,1\}$ and $\boldsymbol{\theta}_m$ represents a vector of model parameters. The corresponding complete data log-likelihood function is,

$$\begin{aligned} \log \mathcal{L}_{com}(\lambda, \boldsymbol{\theta} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) \;=\; & \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \{ M_{ij} \log(\lambda) + (1 - M_{ij}) \log(1-\lambda) + \\ & \qquad M_{ij} \log(\boldsymbol{\pi}_1(i, j; \boldsymbol{\theta}_1)) + (1 - M_{ij}) \log(\boldsymbol{\pi}_0(i, j; \boldsymbol{\theta}_0)) \} \end{aligned}$$

As in the case of the conditional independence assumption, the parameters will be estimated via the EM algorithm. The E-step takes the following form,

$$\xi_{ij} \;=\; \frac{\lambda \boldsymbol{\pi}_1(i, j; \boldsymbol{\theta}_1)}{\lambda \boldsymbol{\pi}_1(i, j; \boldsymbol{\theta}_1) + (1-\lambda) \boldsymbol{\pi}_0(i, j; \boldsymbol{\theta}_0)}$$

whereas the M-Step is as follows,

$$\begin{aligned} \lambda \;&=\; \frac{1}{N_{\mathcal{A}} N_{\mathcal{B}}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} \\ \boldsymbol{\theta}_m \;&=\; \underset{\boldsymbol{\theta}_m^*}{\operatorname{argmax}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij}^m (1 - \xi_{ij})^{1-m} \log(\boldsymbol{\pi}_m(i, j; \boldsymbol{\theta}_m^*)) \end{aligned}$$

As noted by many (e.g., Larsen and Rubin, 2001; Murray, 2016), this second M-step corresponds to the optimization of the weighted log-likelihood function for the log-linear model with a contingency table. In our application, we use the implementation of the log-linear model via fastLink and include all two-way interactions among linkage fields.

Tables A2 and A3 report the match rates and validated turnout rates, respectively. Overall, these estimates presented tend to be somewhat more conservative than the corresponding estimates reported in Tables 4 and 3. Specifically, the resulting matched rates are a few percentage points lower than the matched rates under the conditional independence assumption. As a result, the validated turnout rates are also lower than those obtained under the conditional independence assumption.

		Pre-election	Post-election	Registration rate		
				Voter file		
		fastLink log-linear	fastLink log-linear	all	active	CPS
ANES	Overall	72.11	73.05	80.37	76.57	70.34
		(0.67)	(0.72)			(1.40)
	Internet	72.37	73.45	80.37	76.57	70.34
		(0.79)	(0.85)			(1.40)
	Face-to-face	71.42	72.08	80.22	76.43	70.40
		(1.28)	(1.35)			(1.39)
CCES		60.21	64.41	80.37	76.57	70.34
		(0.19)	(0.21)			(1.39)

Table A2: 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 allowing for dependences across linkage fields ("fastLink" log-linear). 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.

		Pre-election	Post-election	Actual turnout	
				Voter file	Election project
		fastLink log-linear	fastLink log-linear		
ANES	Overall	60.74	62.50	57.55	58.83
		(0.89)	(0.97)		
	Internet	59.49	61.21	57.55	58.83
		(1.01)	(1.14)		
	Face-to-face	64.35	65.94	57.58	58.86
		(1.79)	(1.82)		
CCES		49.56	51.20	57.55	58.83
		(0.31)	(0.37)		

Table A3: Validated Turnout Rates among the Survey Respondents from the 2016 ANES and CCES. The validated turnout rates obtained from the probabilistic model allowing for dependences across linkage fields ("fastLink" log-linear). Those rates 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.

A2.5 Regression Models for Overreporting

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 A4 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 A4, perform a similar analysis but use listwise deletion to deal with missing values. Table A5 repeats this analysis for each interview mode of the ANES.

A2.6 Bivariate Analyses for Overreporting

In this Appendix, we conduct a bivariate analysis of overreporting and present the estimated proportion and odds ratio of overreporting for the values of each covariate of interest. These estimates and their corresponding confidence intervals are obtained using the sample of validated non-voters alone and incorporate the sampling design of each survey. Columns (1) and (3) of Table A6 present the estimated proportion of overreporting, in which non-response is coded as a separate category for each variable to account for missing values. The second set of estimates, shown as in columns (2) and (4) of Table A6, are the odds ratio of overreporting for the different levels of each covariate of interest against a baseline category.

We find that for both the ANES and CCES, those voters who are educated, partisan, and interested in politics are more likely to overreport turnout. In addition, we find that African Americans are more likely to overreport. However, we find no discernible differences across gender and other racial groups. In contrast to our regression analysis, but similarly to other works in the related literature, we find that being married and attending church are associated with a greater likelihood of overreporting. Finally, we find that only for ANES those who are 55 years (or older) are more likely to overreport. Table A7 repeats the same analysis, but the focus is on face-to-face and internet components of the ANES. The results are substantively similar to those presented in Table A6.

A3 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

	ANES				CCES			
	(1)		(2)		(3)		(4)	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Age:								
35 - 44	0.251	0.246	−0.170	0.262	0.140	0.088	0.122	0.100
45 - 54	0.290	0.259	−0.258	0.332	0.100	0.086	0.047	0.097
55+	0.536	0.236	0.471	0.307	0.157	0.082	0.098	0.092
No response	0.167	0.546						
Marital Status:								
Widowed/Divorced	−0.324	0.279	−0.403	0.347	−0.032	0.082	−0.029	0.091
Never married	−0.095	0.186	−0.102	0.247	−0.075	0.074	−0.029	0.085
No response	−0.099	0.207			0.757	0.883		
Education:								
Some College	0.403	0.180	0.300	0.260	0.439	0.075	0.416	0.086
College	0.852	0.230	0.534	0.284	0.645	0.073	0.566	0.082
Post-graduate	0.719	0.247	0.792	0.297	0.764	0.105	0.681	0.118
No response	0.242	0.654						
Gender:								
Male	−0.047	0.130	0.060	0.200	0.268	0.060	0.309	0.067
No response	0.972	0.745						
Race:								
Black	0.789	0.268	0.772	0.404	0.389	0.101	0.418	0.112
Hispanic	−0.215	0.245	−0.464	0.315	−0.295	0.110	−0.178	0.126
Other	−0.510	0.267	−0.645	0.302	−0.591	0.095	−0.527	0.109
No response	0.463	0.980						
Income (in thousands):								
Between 27.5 and 60	0.503	0.212	0.857	0.299	0.499	0.079	0.542	0.085
Between 60 and 100	0.560	0.207	0.585	0.304	0.825	0.089	0.869	0.097
More than 100	0.226	0.264	0.427	0.361	1.033	0.111	1.071	0.120
No response	0.647	0.683			0.643	0.107		
Partisanship:								
Republican	−0.228	0.235	−0.285	0.289	−0.061	0.094	−0.039	0.101
Independent	−0.732	0.190	−0.559	0.234	−0.657	0.074	−0.635	0.080
No response	−1.105	0.433			−1.789	0.128		
Interest in Politics:								
Some	−0.721	0.227	−0.820	0.292	−0.904	0.071	−0.871	0.079
Not very	−1.432	0.238	−1.682	0.337	−1.550	0.086	−1.545	0.096
Not at all	−1.590	0.261	−1.369	0.395	−2.037	0.123	−2.081	0.140
No response	−0.612	0.852			−2.178	0.231		
Church Attendance:								
A few times a year	−0.539	0.191	−0.441	0.262	−0.272	0.089	−0.228	0.099
Rarely/Never	−0.432	0.187	−0.369	0.241	−0.472	0.078	−0.406	0.087
No response	−0.478	0.782			−0.538	0.247		
Ideology:								
Liberal	−0.217	0.368	−0.345	0.423	−0.093	0.135	−0.080	0.146
Moderate	−0.507	0.261	−0.532	0.287	0.086	0.130	0.090	0.141
Conservative	−0.236	0.282	−0.348	0.318	−0.014	0.140	−0.037	0.152
Very conservative	−0.300	0.308	−0.170	0.354	−0.129	0.175	−0.090	0.191
No response	−0.595	0.257			−0.864	0.161		
Intercept	0.981	0.413	1.080	0.574	0.543	0.168	0.451	0.184
Number of observations:	1,390		758		21,976		16,609	

Table A4: Estimated Coefficients for the Weighted Logistic Regression of Overreporting. 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.

	ANES			
	(Face-to-Face)		(Internet)	
	est.	s.e.	est.	s.e.
Age:				
35 - 44	−0.956	0.813	−0.016	0.268
45 - 54	−0.713	0.684	−0.079	0.374
55+	0.364	0.742	0.565	0.343
Marital Status:				
Widowed/Divorced	0.811	0.788	−0.708	0.431
Never married	−0.260	0.650	−0.062	0.281
Education:				
Some College	−0.135	0.649	0.312	0.293
College	0.475	0.617	0.502	0.321
Post-graduate	0.529	0.665	0.722	0.324
Gender:				
Male	0.237	0.410	−0.002	0.232
Race:				
Black	1.406	0.776	0.608	0.457
Hispanic	−0.898	0.531	−0.291	0.411
Other	−1.666	0.774	−0.410	0.315
Income (in thousands):				
Between 27.5 and 60	1.305	0.560	0.843	0.346
Between 60 and 100	1.502	0.483	0.431	0.358
More than 100	1.018	0.853	0.399	0.395
Partisanship:				
Republican	−0.268	0.612	−0.224	0.334
Independent	−0.093	0.610	−0.683	0.257
Interest in Politics:				
Some	−1.017	0.578	−0.887	0.355
Not very	−3.154	0.790	−1.516	0.386
Not at all	−2.241	1.002	−1.393	0.439
Church Attendance:				
A few times a year	−0.220	0.442	−0.597	0.318
Rarely/Never	−0.271	0.488	−0.370	0.298
Ideology:				
Liberal	−0.933	0.855	−0.289	0.490
Moderate	−0.505	0.724	−0.539	0.305
Conservative	−0.877	0.851	−0.332	0.347
Very conservative	−0.707	0.854	−0.174	0.419
Intercept	1.436	1.253	1.131	0.657
Number of observations:	196		562	

Table A5: Estimated Coefficients for the Weighted Logistic Regression of Overreporting by Interview Mode for the ANES. The estimates presented here, and their corresponding standard errors, are obtained from a logistic regression adjusting by the sampling design of the ANES.

	ANES						CCES					
	Proportion			Odds ratio			Proportion			Odds ratio		
		CI			CI			CI			CI	
	est.	2.5%	97.5%	est.	2.5%	97.5%	est.	2.5%	97.5%	est.	2.5%	97.5%
Age:												
18-34	0.36	0.31	0.41	baseline			0.42	0.40	0.44	baseline		
35-44	0.42	0.35	0.50	1.30	0.91	1.86	0.50	0.48	0.53	1.38	1.21	1.57
45-54	0.44	0.35	0.53	1.42	0.94	2.13	0.51	0.48	0.53	1.41	1.23	1.60
55+	0.51	0.45	0.57	1.86	1.35	2.57	0.54	0.52	0.56	1.60	1.42	1.79
No response	0.56	0.40	0.73	2.28	1.09	4.79						
Marital Status:												
Married	0.49	0.44	0.53	baseline			0.53	0.52	0.55	baseline		
Widowed/Divorced	0.38	0.29	0.47	0.64	0.42	0.98	0.46	0.43	0.49	0.74	0.65	0.84
Never married	0.37	0.32	0.42	0.63	0.48	0.82	0.42	0.40	0.44	0.63	0.57	0.70
No response	0.44	0.37	0.52	0.84	0.59	1.18	0.74	0.41	1.06	2.43	0.47	12.63
Education:												
High School or less	0.33	0.28	0.39	baseline			0.36	0.34	0.38	baseline		
Some College	0.44	0.39	0.49	1.60	1.18	2.17	0.51	0.49	0.53	1.83	1.62	2.07
College	0.61	0.54	0.69	3.20	2.14	4.79	0.62	0.60	0.64	2.91	2.59	3.26
Post-graduate	0.63	0.54	0.71	3.37	2.16	5.27	0.72	0.69	0.74	4.46	3.80	5.22
No response	0.50	0.21	0.79	1.99	0.62	6.42						
Gender:												
Female	0.43	0.38	0.47	baseline			0.43	0.42	0.45	baseline		
Male	0.42	0.37	0.46	0.95	0.77	1.18	0.54	0.52	0.56	1.53	1.39	1.68
No response	0.70	0.45	0.95	3.12	0.92	10.54						
Race:												
White	0.42	0.38	0.46	baseline			0.50	0.48	0.51	baseline		
Black	0.58	0.48	0.68	1.92	1.24	2.98	0.55	0.51	0.58	1.22	1.04	1.43
Hispanic	0.35	0.26	0.44	0.75	0.49	1.15	0.38	0.34	0.42	0.62	0.52	0.73
Other	0.33	0.24	0.42	0.69	0.44	1.08	0.39	0.36	0.43	0.66	0.57	0.76
No response	0.72	0.47	0.97	3.58	1.03	12.43						
Income (in thousands):												
Less than 27.5	0.34	0.28	0.39	baseline			0.32	0.30	0.34	baseline		
Between 27.5 and 60	0.45	0.38	0.51	1.61	1.15	2.26	0.49	0.47	0.51	2.05	1.81	2.32
Between 60 and 100	0.49	0.42	0.55	1.87	1.34	2.62	0.60	0.57	0.62	3.21	2.78	3.70
More than 100	0.48	0.40	0.56	1.83	1.21	2.77	0.71	0.68	0.74	5.17	4.35	6.15
No response	0.60	0.42	0.79	3.03	1.31	6.99	0.48	0.44	0.51	1.97	1.65	2.34
Partisanship:												
Democrat	0.53	0.47	0.59	baseline			0.59	0.57	0.61	baseline		
Republican	0.51	0.45	0.58	0.93	0.64	1.36	0.61	0.59	0.64	1.11	0.97	1.27
Independent	0.31	0.27	0.36	0.41	0.30	0.56	0.43	0.41	0.45	0.53	0.47	0.59
No response	0.16	0.05	0.28	0.18	0.07	0.42	0.08	0.07	0.10	0.06	0.05	0.08
Interest in Politics:												
Very	0.71	0.63	0.78	baseline			0.72	0.70	0.74	baseline		
Some	0.48	0.43	0.53	0.39	0.25	0.59	0.46	0.44	0.48	0.33	0.29	0.38
Not very	0.30	0.24	0.36	0.18	0.11	0.28	0.27	0.25	0.29	0.14	0.12	0.17
Not at all	0.23	0.17	0.30	0.13	0.08	0.22	0.13	0.11	0.15	0.06	0.05	0.07
No response	0.54	0.13	0.96	0.50	0.09	2.72	0.11	0.07	0.14	0.05	0.03	0.07
Church Attendance:												
Frequently	0.54	0.48	0.60	baseline			0.61	0.58	0.63	baseline		
A few times a year	0.41	0.35	0.47	0.59	0.43	0.82	0.51	0.49	0.54	0.68	0.59	0.79
Rarely/Never	0.37	0.33	0.42	0.51	0.38	0.70	0.42	0.40	0.43	0.47	0.41	0.53
No response	0.41	0.04	0.77	0.58	0.12	2.88	0.30	0.22	0.38	0.28	0.18	0.42
Ideology:												
Very Liberal	0.57	0.48	0.66	baseline			0.57	0.53	0.62	baseline		
Liberal	0.45	0.32	0.58	0.62	0.34	1.13	0.53	0.50	0.56	0.84	0.67	1.05
Moderate	0.37	0.31	0.44	0.46	0.28	0.73	0.50	0.48	0.52	0.74	0.60	0.91
Conservative	0.47	0.38	0.56	0.67	0.39	1.14	0.57	0.54	0.59	0.97	0.77	1.21
Very conservative	0.58	0.50	0.65	1.04	0.65	1.66	0.57	0.53	0.62	1.00	0.76	1.31
No response	0.32	0.26	0.38	0.36	0.23	0.57	0.11	0.09	0.12	0.09	0.07	0.11

Table A6: Estimated Proportion and Odds Ratio Across Different Response Levels of Standard Predictors of Overreporting. The estimates presented here, and their corresponding confidence intervals, are obtained adjusting by the sampling design of the ANES and the CCES.

	ANES											
	Face-to-Face						Internet					
	Proportion			Odds ratio			Proportion			Odds ratio		
		CI			CI			CI			CI	
	est.	2.5%	97.5%	est.	2.5%	97.5%	est.	2.5%	97.5%	est.	2.5%	97.5%
Age:												
18-34	0.35	0.25	0.45	baseline			0.36	0.31	0.42	baseline		
35-44	0.44	0.28	0.60	1.43	0.76	2.70	0.42	0.33	0.51	1.25	0.81	1.93
45-54	0.35	0.17	0.54	1.00	0.42	2.38	0.47	0.37	0.57	1.54	0.96	2.48
55+	0.51	0.42	0.60	1.92	1.04	3.57	0.51	0.44	0.58	1.83	1.25	2.67
No response	0.53	0.24	0.82	2.06	0.55	7.64	0.58	0.38	0.78	2.40	0.99	5.79
Marital Status:												
Married	0.46	0.38	0.54	baseline			0.50	0.44	0.55	baseline		
Widowed/Divorced	0.44	0.32	0.55	0.92	0.49	1.70	0.35	0.23	0.47	0.55	0.31	0.95
Never married	0.33	0.24	0.43	0.59	0.38	0.93	0.39	0.33	0.45	0.64	0.47	0.88
No response	0.48	0.35	0.62	1.11	0.64	1.92	0.43	0.34	0.52	0.75	0.49	1.15
Education:												
High School or less	0.29	0.19	0.38	baseline			0.35	0.28	0.41	baseline		
Some College	0.43	0.33	0.53	1.89	0.98	3.65	0.45	0.39	0.50	1.51	1.07	2.14
College	0.62	0.50	0.75	4.12	2.12	7.98	0.61	0.52	0.70	2.94	1.80	4.83
Post-graduate	0.77	0.60	0.94	8.30	2.79	24.71	0.59	0.50	0.69	2.71	1.66	4.44
No response	0.12	−0.13	0.36	0.32	0.03	3.33	0.60	0.28	0.91	2.75	0.71	10.69
Gender:												
Female	0.38	0.29	0.48	baseline			0.44	0.39	0.49	baseline		
Male	0.43	0.35	0.51	1.20	0.76	1.89	0.41	0.36	0.46	0.88	0.68	1.13
No response				3.12	0.92	10.54	0.70	0.45	0.95	2.92	0.86	9.91
Race:												
White	0.40	0.31	0.48	baseline			0.42	0.38	0.47	baseline		
Black	0.62	0.43	0.82	2.53	1.09	5.90	0.57	0.45	0.68	1.78	1.07	2.98
Hispanic	0.34	0.24	0.45	0.80	0.47	1.38	0.36	0.24	0.48	0.77	0.44	1.33
Other	0.27	0.11	0.44	0.58	0.23	1.47	0.35	0.24	0.47	0.75	0.44	1.26
No response	0.68	0.16	1.21	3.31	0.27	40.50	0.73	0.45	1.01	3.64	0.87	15.30
Income (in thousands):												
Less than 27.5	0.27	0.19	0.36	baseline			0.36	0.29	0.42	baseline		
Between 27.5 and 60	0.43	0.28	0.58	1.97	1.00	3.89	0.46	0.39	0.53	1.51	1.03	2.23
Between 60 and 100	0.56	0.46	0.67	3.41	1.84	6.30	0.46	0.38	0.54	1.52	1.02	2.27
More than 100	0.53	0.32	0.73	2.95	1.16	7.50	0.47	0.38	0.56	1.61	1.01	2.57
No response	0.44	0.10	0.78	2.08	0.53	8.20	0.71	0.53	0.88	4.32	1.71	10.93
Partisanship:												
Democrat	0.54	0.40	0.68	baseline			0.53	0.46	0.60	baseline		
Republican	0.44	0.30	0.57	0.67	0.26	1.70	0.54	0.47	0.61	1.04	0.69	1.57
Independent	0.37	0.27	0.47	0.50	0.28	0.91	0.30	0.24	0.35	0.37	0.26	0.54
No response	0.10	0.01	0.20	0.10	0.03	0.29	0.47	0.08	0.86	0.78	0.16	3.88
Interest in Politics:												
Very	0.68	0.56	0.80	baseline			0.71	0.63	0.80	baseline		
Some	0.51	0.40	0.61	0.48	0.24	0.97	0.47	0.41	0.53	0.36	0.21	0.59
Not very	0.23	0.13	0.33	0.14	0.06	0.32	0.32	0.25	0.39	0.19	0.11	0.31
Not at all	0.17	0.10	0.23	0.09	0.04	0.19	0.26	0.17	0.34	0.14	0.07	0.26
No response	0.64	0.01	1.28	0.85	0.06	12.28	0.47	−0.02	0.97	0.36	0.05	2.84
Church Attendance:												
Frequently	0.49	0.41	0.57	baseline			0.56	0.48	0.64	baseline		
A few times a year	0.46	0.35	0.57	0.87	0.51	1.51	0.39	0.32	0.46	0.50	0.34	0.75
Rarely/Never	0.33	0.24	0.41	0.50	0.32	0.80	0.39	0.34	0.44	0.51	0.35	0.74
No response							0.54	0.09	0.99	0.93	0.14	6.34
Ideology:												
Very Liberal	0.70	0.55	0.85	baseline			0.52	0.42	0.63	baseline		
Liberal	0.50	0.33	0.67	0.42	0.16	1.13	0.43	0.27	0.60	0.70	0.33	1.45
Moderate	0.40	0.29	0.51	0.28	0.12	0.67	0.37	0.29	0.44	0.52	0.30	0.91
Conservative	0.46	0.26	0.66	0.36	0.12	1.05	0.47	0.38	0.57	0.82	0.45	1.49
Very conservative	0.49	0.36	0.63	0.41	0.17	1.00	0.61	0.52	0.70	1.42	0.81	2.46
No response	0.22	0.12	0.33	0.12	0.06	0.27	0.36	0.29	0.43	0.51	0.30	0.86

Table A7: Estimated Proportion and Odds Ratio Across Different Response Levels of Standard Predictors of Overreporting. The estimates presented here, and their corresponding confidence intervals, are obtained adjusting by the sampling design of each component of the ANES.

		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 A8: 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.

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 A8 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 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 A4 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

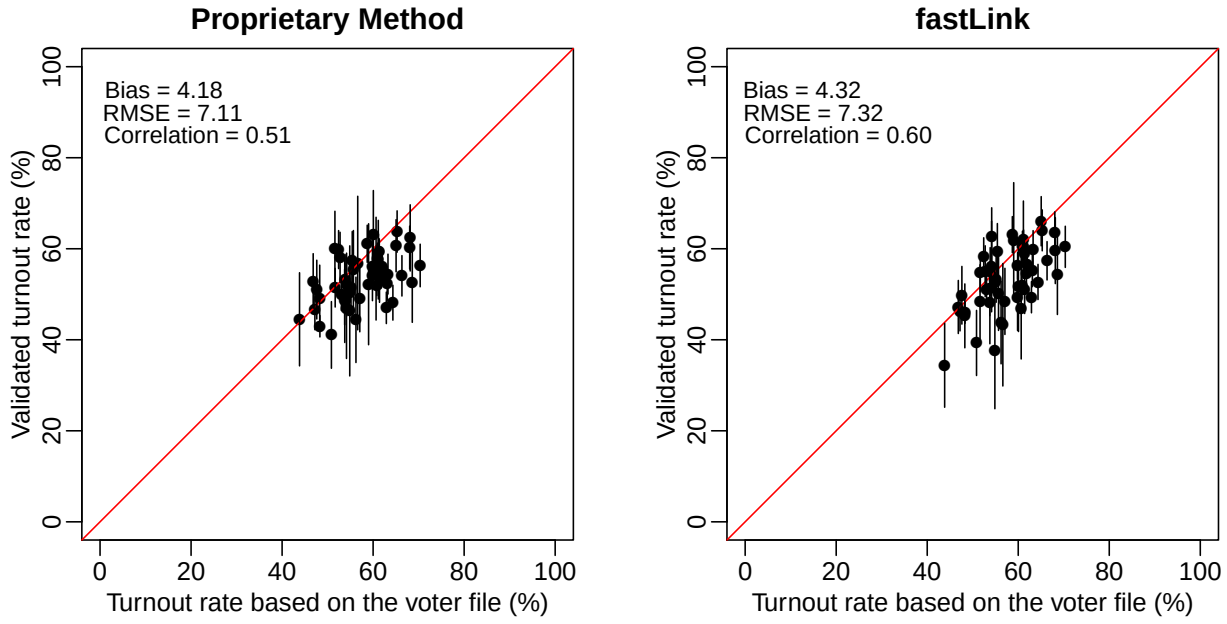


Figure A4: Comparison of the Validated Turnout Rates against the Actual Turnout Rates at the State-level. We evaluate the performance of the proprietary method (left plot) and fastLink (right plot) by plotting the resulting state-level validated turnout rate (vertical axis) against the actual turnout rate based on the voter file (horizontal axis).

the vertical axis represents the validated turnout 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 algorithms 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 overreporting. The estimated coefficients and their standard errors are given in Table A9 of Appendix A4, which in addition contains a complete description how the estimation was conducted.

Figure A5 presents the predicted probabilities for four possible matching statuses based on the fitted

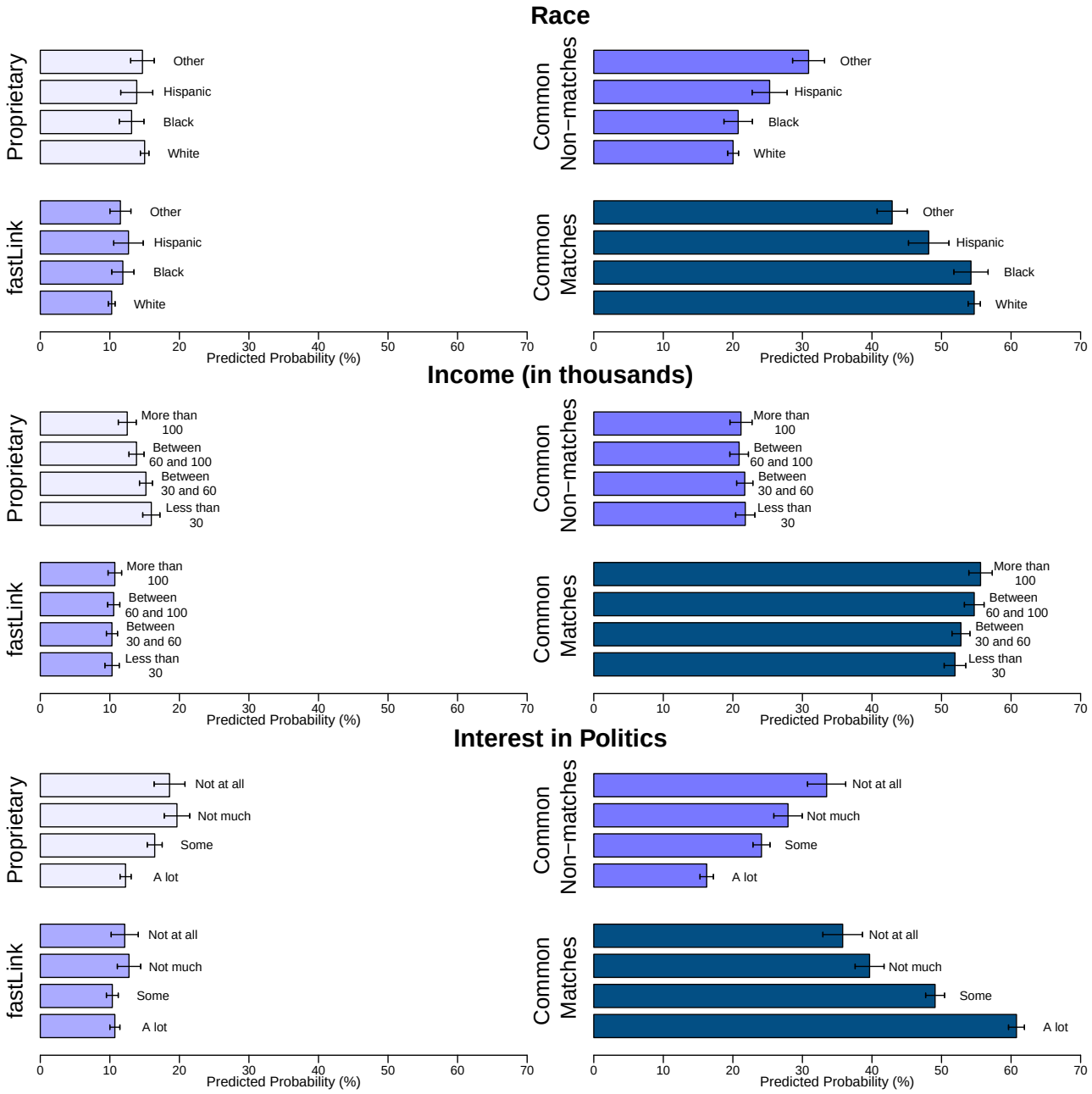


Figure A5: Predicted Probabilities for the Matching Status of the Different Vote Validation Exercises across Covariates. The results are based on the weighted multinomial logistic regression, where the outcome variable takes four values, indicating different matching status for each CCES respondent: 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.

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 individuals.

A4 Additional Details of the 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 A5 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.

A5 Address Merge for the 2016 ANES Study

Although names and date of birth often contain measurement error, the ANES has validated all the addresses that were selected in the sampling process. We take advantage of this by merging 4,271 sampled addresses of the 2016 ANES pre-election survey with more than 110 million addresses in the voter file. In particular, we conduct the merge using the house number, street name, and apartment number as linkage fields after blocking on the zip code. We use two levels agreement for street name based on the Jaro-Winkler distance and 0.94 as the threshold. For the remaining variables, we use a binary comparison of whether or not two records have an identical value.

To estimate the estimate the turnout rate from the address merge, we use the following approximation,

$$\frac{\sum_{j=1}^{J} w_j \zeta_j T_j}{\sum_{j=1}^{J} v_j w_j} \tag{A1}$$

where J represents the total number of sampled ANES households, T_j is the total number of registered voters who live in address j and voted in the 2016 election, w_j is the sampling weight for address j provided by the ANES, ζ_j is the estimated match probability for address j provided by fastLink (as defined in equation 4), and v_j is the average number of voting-age individuals in the census block group where address j is located. Note that out of the 4,271 addresses we were able to match 3,814 addresses, for the remaining 457 addresses we obtain the average number of voting-age individuals from addresses that are matched based on zip code and street name but differ in the house number.

Table A10 reports the turnout rates based on this address merge. We find that these estimated turnout rates are close to the actual turnout rates. Given that the denominator presented in equation A5 does not exclude non-citizens and other ineligible individuals of voting age, it is not surprising that our estimates from the address merge are slightly lower than the actual turnout rates. Like our respondent-level results, the estimated turnout rates for the pre-election and post-election samples are within one percentage point difference from each other, suggesting that attrition is adjusted properly by sampling weights.

We also find that the estimated registration rates are close to the actual registration rates based on the voter file and the CPS. Again, the pre-election and post-election estimates are similar to each other (see

Table A11).

References

American National Election Studies (2017). User's guide and codebook for the ANES 2016 time series study. Tech. rep., University of Michigan and Stanford University, Ann Arbor, MI and Palo Alto, CA.

Ansolabehere, S., Schaffner, B., and Luks, S. (2017). Guide to the 2016 cooperative congressional election survey. Tech. rep., Harvard University. Data Release No. 2.

DeBell, M., Amsbary, M., Meldener, V., Brock, S., and Maisel, N. (2016). Methodology report for the anes 2016 time series study. Tech. rep., Stanford University and the University of Michigan., Ann Arbor, MI and Palo Alto, CA.

Enamorado, T., Fifield, B., and Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review* **113**, 2, 353–371.

Larsen, M. D. and Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association* **96**, 453, 32–41.

Murray, J. S. (2016). Probabilistic record linkage and deduplication after indexing, blocking, and filtering. *Journal of Privacy and Confidentiality* **7**, 1, 3–24.

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Survey Methodology.* 31–38.

Winkler, W. E. (1989). Near automatic weight computation in the fellegi-sunter model of record linkage. Tech. rep., Proceedings of the Census Bureau Annual Research Conference.

Winkler, W. E. (1993). Improved decision rules in the fellegi-sunter model of record linkage. In Proceedings of Survey Research Methods Section, American Statistical Association.

	Common Non-Matches		fastLink only		Common Matches	
	est.	s.e.	est.	s.e.	est.	s.e.
Age:						
35-44	−0.295	0.045	−0.221	0.056	0.070	0.043
45-54	−0.533	0.050	−0.138	0.059	0.167	0.045
55+	−0.791	0.045	−0.211	0.053	0.482	0.040
Marital Status:						
Widowed/Divorced	0.019	0.046	0.096	0.053	−0.002	0.038
Never married	0.045	0.040	0.232	0.047	0.210	0.036
No response	0.995	0.558	−0.069	0.809	0.007	0.564
Education:						
Some College	−0.176	0.040	0.182	0.047	0.220	0.035
College	−0.022	0.040	0.240	0.047	0.248	0.035
Post-graduate	0.193	0.064	0.387	0.072	0.252	0.055
Gender:						
Male	0.258	0.032	0.232	0.037	0.147	0.028
Race:						
Black	0.182	0.050	0.268	0.058	0.100	0.045
Hispanic	0.351	0.056	0.278	0.067	−0.099	0.054
Other	0.519	0.053	0.141	0.065	−0.281	0.052
Income (in thousands):						
Between 30 and 60	0.028	0.040	0.043	0.049	0.060	0.036
Between 60 and 100	0.096	0.047	0.179	0.056	0.199	0.042
More than 100	0.235	0.057	0.316	0.066	0.319	0.050
No response	0.208	0.058	0.445	0.065	0.151	0.051
Partisanship:						
Republican	−0.259	0.050	−0.322	0.059	−0.160	0.043
Independent	−0.015	0.040	−0.093	0.047	−0.269	0.036
No response	0.198	0.061	−0.057	0.077	−0.789	0.065
Interest in Politics:						
Some	0.153	0.038	−0.305	0.044	−0.551	0.032
Not very	0.139	0.048	−0.274	0.057	−0.957	0.044
Not at all	0.387	0.058	−0.260	0.073	−1.020	0.059
No response	0.268	0.130	−0.298	0.169	−1.095	0.147
Church Attendance:						
A few times a year	−0.175	0.045	−0.215	0.052	−0.162	0.039
Rarely/Never	−0.236	0.040	−0.210	0.046	−0.133	0.035
No response	0.158	0.127	−0.504	0.179	−0.192	0.128
Ideology:						
Liberal	0.275	0.076	0.155	0.085	0.078	0.064
Moderate	0.313	0.072	0.173	0.081	−0.015	0.061
Conservative	0.429	0.077	0.244	0.088	0.027	0.066
Very conservative	0.249	0.094	0.018	0.109	0.019	0.079
No response	0.372	0.087	0.190	0.103	−0.267	0.080
Intercept	0.230	0.091	−0.441	0.105	1.299	0.079

Table A9: 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.

		Validated Turnout (fastLink)		Actual turnout	
		Pre-election sample	Post-election sample	Voter file	Election project
ANES	Overall	54.01 (1.02)	54.74 (1.14)	57.55	58.83
	Internet	54.00 (0.81)	54.50 (0.85)	57.55	58.83
	Face-to-face	54.06 (2.99)	55.31 (3.34)	57.58	58.86

Table A10: Validated Turnout Rates among the Residents of the Addresses from the ANES 2016 Study. The validated turnout rates obtained from the probabilistic model alone ("fastLink") 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.

		Match Rates (fastLink)		Registration rate		
		Pre-election sample	Post-election sample	Voter file		
				all	active	CPS
ANES	Overall	67.97 (0.84)	68.03 (0.94)	80.37	76.57	70.34 (1.40)
	Internet	68.13 (0.81)	68.05 (0.90)	80.37	76.57	70.34 (1.40)
	Face-to-face	67.55 (2.16)	67.99 (2.37)	80.22	76.43	70.40 (1.39)

Table A11: Match Rates among the Residents of the Addresses from the ANES 2016 Study. 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"). 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).