

Towards Credible Vote Validation

Matthew DeBell¹

¹ *Stanford University*

Abstract

Vote validation (VV) is becoming standard practice in academic election surveys as a solution to frequent over-reporting of voter turnout by survey participants, and VV results are usually presumed accurate. Yet evidence for VV's accuracy is limited. This article investigates these limits to help calibrate our confidence in the process by which VV data are generated and to improve that process. Concerns with VV are serious: voter files and survey data contain errors; state-of-the-art matching procedures make errors; resulting estimated match probabilities are misleading; part of the process (clerical review) has lacked standards; and evidence purported to show VV's accuracy is consistent with overall accuracy similar to self-reports. Considering these concerns, I present a transparent and reliable clerical review procedure to evaluate pairs of matched records, I present an interpretive framework for evaluating accuracy, and I argue for skepticism about VV's accuracy so that researchers make better-informed choices about when and how to use VV data. Self-reports have better sensitivity (counting nearly all voters) while validation has better specificity (excluding more non-voters), and the two measures are subject to different kinds of error, so rather than preferring one automatically, researchers should consider which is more accurate for a given research question and, where possible, compare results from both measures.

Keywords: vote validation, record linkage, validated turnout, turnout misreporting



“...who these ‘not-founds’ might be has troubled users of the validation data. How do we know that these are not simply difficult to find voters?”

—ANES Director Santa Traugott (1989)

Understanding who votes is fundamental to the study of voting behavior, elections, and democracy. Yet surveys badly overestimate voter turnout and have done so for decades (DeBell et al., 2020; Parry & Crossley, 1950; Traugott et al., 1992). One important source of this error is thought to be misreporting by non-voters who falsely say they voted. These misreports adversely affect nearly all voting behavior surveys by distorting our picture of the electorate. In search of better data to correctly distinguish voters from non-voters, researchers have attempted to verify or falsify survey reports by linking respondents to administrative records of their registration and turnout—a process called “vote validation” (VV).

The American National Election Studies (ANES) conducted VV for most of its studies between 1964 and 1988 but abandoned the practice in the early ‘90s because it was expensive and unreliable at linking survey respondents to voting records (Rosenstone & Legee, 1994). More recently, changes in VV’s accessibility have led to its renaissance. Miller et al. (2021, p. 344) recommend that “all election studies should seek to validate respondent reports of voting,” and VV data are now routinely included in major surveys in the United States, including the ANES, Cooperative Election Study (CES), AP VoteCast, and established survey panels such as YouGov and NORC AmeriSpeak.

The renaissance of vote validation is based on the beguiling conclusion that it is a reliable procedure (Ansolabehere & Hersh, 2012; Enamorado & Imai, 2019) that can “eliminate the overreporting problem” (Nyhan et al., 2017, p. 758). The prevalence of this view is evident in the commonplace use of VV data, which have been the basis for highly cited work on important subjects including who votes (Ansolabehere & Hersh, 2012), intention to vote (Achen & Blais, 2016), why people vote (Bond et al., 2012), minority turnout in majority-minority districts (Fraga, 2016), the heritability of political behavior (Fowler et al., 2008), and, of course, overreporting itself.

I argue that VV has serious shortcomings: it has frequent errors, does not eliminate the overreporting problem, and should not be automatically preferred

Direct correspondence to

Matthew DeBell, Institute for Research in the Social Sciences, Stanford University,
Stanford, CA, USA
e-Mail: debell@stanford.edu

to self-reports. But it has value in being differently biased than self-reports, and it can be improved. The argument proceeds in four sections. First, rather than establishing that VV is accurate, the literature gives cause for concern. Second, endemic data errors afflict both official records and survey data, making it difficult or impossible to correctly link some records. This tempers the confidence we can have in VV results. Third, limitations of probabilistic record linkage—the leading VV method—include estimated “match probabilities” that should not be treated as probabilities and are too often facially invalid, and linkage procedures that have lacked a transparent or reliable method for the required clerical review; I present one. Finally, to evaluate VV’s accuracy and fitness for use, scholars need an interpretive framework that considers all relevant aspects of error. The live controversy in VV literature concerns VV’s accuracy. I apply a new interpretive framework, revealing that VV is not necessarily more accurate than self-reports and emphasizing a tradeoff between strengths (VV is better at excluding non-voters) and weaknesses (VV is worse at including all voters). Awareness of these issues will equip researchers to improve VV and treat it with due skepticism.

Cause for Concern

The ANES first performed VV in connection with its 1964 study and continued these efforts until the early 1990s. In each sampled locality, interviewers visited elections offices and manually reviewed public records of registered voters to record the registration and turnout status of ANES participants. The exercise revealed reasons to be concerned about validation failures, including the one expressed by Santa Traugott in this article’s epigraph. The likelihood of finding voters varied with the quality of government files and also varied with respondents’ characteristics such as race, home tenure, socioeconomic status, and gender, which distorted the correlates of turnout (Presser et al., 1990). Concerns about reliability led to a revalidation effort in which ANES respondents from 1988 had their turnout status checked in 1989 and again in 1991 (Traugott et al., 1992). It did not go well: only 87% of results were reproduced on the second try. This was the death knell of VV efforts in the 1990s, as the results produced “at best, a marginal improvement in the reliability of the measure of voter turnout”, and the expense was not justifiable (Rosenstone & Leege, 1994, p. 697).

Commercial Voter Files

The emergence of commercially aggregated voter files in the 2008 election cycle made VV more accessible and affordable, heralding the renaissance of validation efforts. Ansolabehere and Hersh (2012) reported the first study using a com-

mercial file for nationwide all-electronic VV, hiring Catalist to match records to the 2008 CES. They defended the accuracy of VV, stressing three reasons for their confidence in the matches: Catalist's description of its data processing, testing of Catalist's matching procedure, and confirmatory evidence of successful matching in their analysis of CES data.

Catalist aggregated nearly all voter records nationwide into one electronic file. Catalist's data processing retained inactive records, so turnout histories were not purged if a voter's registration lapsed (as is the case with many government files), and included other valuable processing steps such as de-duplicating in-state records and tracking movers using the National Change of Address database. However, they also imputed gender (for 28% of records) and birth date when these were missing, which may reduce matching accuracy when imputation is incorrect. When Ansolabehere and Hersh (2012) sampled directly from official voting records and then matched them to Catalist's database, they found many aspects of close correspondence but also found that Catalist incorrectly identified the voting status of 6% of voters and 4% of non-voters. Among survey respondents, self-reported and validated characteristics such as race and party identification were incorrect 5% to 7% of the time among matched cases. Confirmatory evidence included analysis of misreporting respondents that compared matched and unmatched cases, identified consistency across election years, and controlled for an index of the quality of official records. It was beyond the scope of their study to consider whether the quality of official records might have substantial components that were not captured by the quality index or to consider how consistent errors could produce consistent matching results over time. They also did not directly examine errors on voter files or on survey datasets, which I address below in detail. Thus, while their study broke important new ground and gave evidence about VV's strengths, it should not be read as showing that validation errors are ignorable.

ANES Validation 2008

Working with ANES data, Berent et al. (2016) challenged the notion that VV gives more accurate results than survey self-reports. They attempted validation using data obtained directly from elections officials in six states and found matches in the government records for just 46% to 77% of survey respondents, depending on the matching criteria used. The lower end of this range was produced using deterministic matching rules that called for an exact match between the two datasets for the name, address (after standardizing to USPS format), and date of birth. It is not believable that only 46% of respondents are included on voter files, because a much higher percentage of the population is registered to vote. The low match rate reflects commonplace failures to match when using

strict criteria. This showed that exact matching between records is not a viable method due to discrepancies, missingness, or errors in one or both data sources.

Berent et al. (2016) found that among the cases they successfully matched to registered voter records, only 6% misreported turnout by saying they voted when their official record said they did not. Based on frequent matching failures and errors in the data, they remained agnostic about the turnout status of cases that were not matched to registration records. This led to the conclusion that the apparent accuracy of validated turnout estimates “is most likely an illusion” (Berent et al., 2016, p. 599) produced by matching failures and that misreporting of turnout is a relatively small factor in the “turnout gap” (i.e., the excess of self-reported voters in sample surveys relative to the proportion of voters in the population).

Probabilistic Matching

Berent et al. (2016) noted the possibility that different matching procedures could overcome the problems they described. A solution was proposed by Enamorado et al. (2019), who developed open-source matching software, *fastLink*, that substantially improves the transparency, speed, and reproducibility of the matching process compared to deterministic matching and proprietary matching by commercial firms. It also improves on deterministic matching by quantifying the uncertainty about match quality caused by discrepancies and missing data. They partnered with ANES to use *fastLink* to produce validated turnout data for the ANES 2016 Time Series Study (Enamorado et al., 2018), and the ANES used *fastLink* and other probabilistic matching to conduct VV for 2020 (ANES, 2022). Probabilistic matching is now the state of the art for VV, and *fastLink*, as the 2021 winner of the Society for Political Methodology’s Best Statistical Software Award, is the leader in probabilistic matching.

Record linkage relies on comparing a record from one file to all records in a reference file and, as such, the computational demands are a function of the product of the file sizes. To make this computationally feasible, “blocking” is used, in which comparisons are limited to records with certain characteristics. Enamorado and Imai (2019) blocked on gender and state, and within these blocks they matched probabilistically on first name, last name, street number, street name, age, and zip code. Numerous potential matches are generated for most cases, and *fastLink* calculates a “match probability” for each that is said to represent the probability of correct linkage for each survey respondent. As such, turnout analysis can weight cases by their match probability to generate estimates that account for potential matching errors. That is, the product of match probability and validated turnout should constitute the posterior probability of voting. Conceptually, this is an elegant solution to the problem of uncertainty in

matching; I return to this issue in the Limitations of Probabilistic Record Linkage section.

The fastLink software was tested in several ways by Enamorado et al. (2019). When they attempted record linkage between the 2014 and 2015 L2, Inc., voter files, setting a match probability threshold of .85, they obtained a match rate of 94%, which they call “very high” (p. 366). It is higher than the ANES achieved in its 1989-1991 study cited earlier (Traugott et al., 1992), but a 6% failure rate is the same as one estimate of the overreporting rate (Jackman & Spahn, 2019). They also conducted a simulation study matching women on the California voter file with errors introduced to the data. These tests are helpful and more could be done to establish that fastLink delivers accurate matching results with national datasets that have errors realistically representing the errors in survey data. Realistic errors include typographical errors, missing data, switched gender, discrepant addresses, and the other kinds of errors appearing in data files described earlier. We do not know how frequently fastLink fails to identify credible matches in a nationwide voter file under these realistic conditions.

Part of Enamorado and Imai’s (2019) process was an “optional” clerical review that made the ANES result, but not the CES result, appear more accurate. They did not detail any coding criteria for this review. They discarded 8.7% of fastLink’s ANES matches as being not credible, after which the validated turnout estimate was within 1 percentage point, and within sampling error, of the actual turnout rate. That seems like a bullseye result, but importantly, this is a finding that substantially all of the ANES turnout gap is attributable to overreporting, so nonresponse bias is, according to this result, practically nonexistent. The seeming bullseye actually reveals an important error because we know the ANES has non-response bias related to voter turnout. We know this from three congruent research findings. First, members of the ANES sample who readily cooperate differ from those who require more sustained efforts to interview, with the “hard to get” group (those who require more contact attempts, escalated monetary incentives, or refusal conversion efforts) having lower turnout, as demonstrated with data from the 1980s (Peress, 2010) and 2016 (DeBell et al., 2018); this is evidence that ANES response propensity is related to voter turnout. Other voter surveys confirm this result (e.g., Dahlgaard et al., 2019; Tourangeau et al., 2010). Second, ANES 2016 non-respondents, 657 of whom were finally interviewed in a non-response follow-up study long after the ANES was over, were found to have lower self-reported turnout rates than ANES respondents had, with a difference of 10 percentage points (DeBell et al., 2018), giving direct evidence that ANES non-response is related to turnout. Third, VV on the full ANES sample of responding and non-responding households alike also shows lower turnout in ANES non-responding households than among those that respond (DeBell, Maisel, et al., 2022). ANES is typical in having this bias, which is found in election studies worldwide (Selb & Munzert, 2013). To be credible when non-response bias is known to exist, any VV result must estimate validated turnout

that is higher than the true (official) turnout rate to account for some degree of nonresponse bias.

The concerns raised above show that VV’s accuracy is limited not only by matching methods but also by the quality of the underlying data. Many validation errors arise from errors in both voter files and survey datasets. These errors can produce an appearance of topline accuracy even when individual VV results are incorrect. To understand the limits of VV, it is therefore necessary to examine data errors that shape validation outcomes.

Endemic Data Errors

The literature employing VV has discounted several sources of error that reduce matching accuracy. This section details these sources of error and their likely magnitude. There are two domains: incorrect information in the voter files, including mislistings and erroneous turnout records, and incorrect information in the survey datasets, including measurement error and selection error. The errors are summarized in Table 1.

Table 1

Dataset Errors and Their Estimated Magnitude

Error	Percent of records affected	Result for correct match probability
On voter files		
Duplicate registrations	2	Lower
Incorrect address for registered voter	8	Much lower
Incorrect address for non-registered	4	Much lower
Omitted from file	14	Zero
Missing date of birth	14	Lower
True voter's turnout not recorded	4	False negative conditional on record match
On survey datasets		
Date of birth error	3.3	Lower
Gender error	1.4	Zero (if blocking)
Name error	3.7	Lower
Panel respondent substitution	3.3	Near zero

Note. Sources: Ansolabehere and Hersh (2010); DeBell, Amsbary, et al. (2022); Enamorado and Imai (2019); Jackman and Spahn (2021); Pew (2012).

Incorrect information in the voter files

Datasets used for voter validation—government files or commercially aggregated data—contain many errors. In data from 2012, about 2% of registrations were duplicates (Pew Center on the States, 2012), often due to moving. About 12% of the U.S. population was living somewhere other than their recorded address (Jackman & Spahn, 2021), and about two-thirds of those mislisted cases (amounting to 8% of the population) were registered to vote. In data from a vendor used for the ANES 2020 VV, 14% of the voting-eligible population was missing (see Online Appendix 1, footnote 1). Ansolabehere and Hersh's (2010) review of Catalist data found that 14% were missing date of birth.

While the errors above reduce the likelihood of successful record matching, there are also discrepancies in the official turnout data themselves, such that the voter files understate turnout compared to official results or self-reports. To understand these discrepancies, consider how a vote means something different for a voter, a registrar, and an election result. For in-person voting, a voter reasonably thinks she voted if she operates a voting machine; a registrar records the voter as having voted if a poll worker duly records the event in the rolls. Separately, election administrators add a vote to the tally if the voting machine functions properly and the ballot is marked for the office. Thus, the three outcomes can differ due to errors or undervotes. And differ they do. In 2016 the number of voters on the L2 voter file was nearly 6 million fewer than the number of votes counted (131 million compared to 136.7), omitting about 4% of voters (Enamorado & Imai, 2019). These errors in the voter files make for downward-biased turnout estimates.

Commercial voter files give inconsistent results that reflect errors. In a Pew Research Center study of five nationwide voter files, “match” rates by firms linking to the same survey dataset ranged from 50% to 79% while validated turnout rates for the 2016 election ranged from 71% to 86% (Igielnik et al., 2018). Among the cases matched by all five firms, the turnout rates ranged from 81% to 88% for the same matched individuals. The higher turnout rate among cases that were matched on all files suggests a correlation between matching status and voting behavior, which would be a further source of bias in VV. This demonstrates that the commercial voter files have different content, are collectively unreliable, and that errors in vendor databases or matching lead to vote validation errors that may equal or exceed survey self-reporting errors.

Incorrect Information in Surveys

All survey data are subject to misreporting, but rates of misreporting are infrequently known. A close review and data cleaning effort involving follow-up contacts to resolve inconsistencies in the 2020 ANES estimated that 3.3% of respondents have recorded date of birth incorrectly, 1.4% have recorded gen-

der incorrectly, and that different names were given at different times for the same person in 3.7% of the ANES sample (DeBell, Amsbary, et al., 2022). Thus, ANES has errors in variables critical for record linkage to external databases. ANES spares little expense to maximize data quality, so these numbers reflect a potential best-case condition for the accuracy of matching variables in Internet survey data.

In addition to reporting errors, surveys are subject to selection errors in which the wrong person answers the questionnaire under someone else's name. This can defeat validation attempts, particularly in panel studies such as ANES or CES/YouGov, where background information is collected in one wave and turnout is collected later. For example, in each presidential election year the ANES is a two-wave panel study, with a pre-election interview and a post-election interview. The respondent reports their gender, date of birth, and the name under which they are registered to vote once, during the pre-election interview. Turnout is usually reported in the post-election interview. Similarly, the CES asks background information in a profile survey separate from the questionnaire in which turnout is reported. If another person substitutes for the sampled individual in one of these interviews, the dataset will show one's name, gender, and date of birth combined with the other's turnout report. Such substitution is strictly not allowed, yet sometimes happens, particularly in self-administered internet studies when a respondent becomes confused or declines to follow instructions. The ANES carefully examined this for the first time in the 2020 study. Cases were flagged for discrepancies between the screener, pre-election, and post-election data, and a sample was selected for follow-up contacts to verify the data. Substitution was confirmed in at least 1.7% of cases and estimated to have occurred in 3.3% of cases (DeBell, Amsbary, et al., 2022). When substitution occurs, VV has little to no chance of correct record linkage for the person who self-reported their turnout status.

Table 1 summarizes the errors discussed in this section. Duplicate registrations, missing or incorrect date of birth, name errors, and incorrect addresses all reduce the probability of a correct match and increase the probability of a matching error. While these errors are not mutually exclusive, they sum to about 35% of records as an upper bound. Respondent substitution in a panel study reduces the match probability nearly to zero, as does omission of a person from the commercial voter file, and a gender error on the survey file makes a correct fastLink match impossible when following Enamorado and Imai's (2019) blocking procedure; these errors that make a correct match practically impossible sum to about 18%. When a true voter's turnout is not recorded, this will result in a false negative result for any matches to that record, which will bias the validated turnout rate downward by about 4 points. Collectively, the errors potentially result in a match failure or error rate well into the double digits and in at least a 4-point downward bias in turnout estimates when matches are made.

These pervasive errors limit the accuracy of record linkage. The next section turns to limitations of probabilistic record linkage procedures themselves.

Limitations of Probabilistic Record Linkage

Oscar Madison and Felix Ungar share an apartment.¹ Suppose Oscar is not a registered voter but answers a survey. If Felix is registered, would VV mistakenly attribute his turnout history to Oscar? Probabilistic matching without clerical review makes this error and others surprisingly likely. This section elaborates difficulties with probabilistic matching, including misleading match probabilities, ambiguous clerical review standards, the need for more realistic software testing, and the likelihood that many non-matched cases may be voters. An improved clerical review procedure overcomes some of these limitations.

Facially Invalid Probabilities

fastLink is vulnerable to several types of errors that result in match probabilities lacking face validity. Table 2 illustrates such errors by showing several pairings of simulated records with their match probability scores. To protect respondent privacy, examples with actual respondents cannot be shown, so examples with simulated data are given here to illustrate errors observed in ANES VV. For each numbered pair, the first row describes a hypothetical survey respondent, and the second describes a potential match and gives fastLink's match score. To generate these results, a simulated survey sample file of 100 cases was matched to the Vermont voter file, with the voter file seeded with additional records, and matching on first name with suffix, last name, house number, street name, zip code, and birth year as separate fields (detailed in Online Appendix 2).

Several living scenarios can generate false but high-confidence matches. Parents and their adult children who live together can match, as illustrated in pair 1: Ricky Ricardo, Sr, born in 1917, matches with a probability > 99.9% to his apparent son, Enrique Ricardo, Jr, born in 1953, despite opposing Jr and Sr suffixes, a large age difference, and Ricky's use of a nickname.

This can also happen in the more common case where parents and their children have different names and no suffix, as in pair 2: Tim Cratchit, born in 2000, matches the 25-years-older Bob Cratchit with a probability of > 99.9%. Same-sex couples or roommates can also be matched with high but false confidence even when their ages differ and their names have no resemblance, such as Ellen Ripley and Sarah Connor (92.0%, pair 3). If two people with the same first name but different last names happen to live together or to be associated with the same

¹ These are the main characters in *The Odd Couple*, Neil Simon's 1965 play and subsequent film and TV series.

address because one is a recent tenant, they can also match (pair 4, where Elizabeth Lemon, born in 1974, is taken for the 33-years-older Elizabeth Bennett at 99.2%).

Table 2

Troublesome Match Probabilities From fastLink (Simulated Data)

Pair	Name	Address	ZIP	Birth year	Match prob.
1	Ricky Ricardo, Sr	1270 Montgomery Ave	05447	1917	> 99.9
	Enrique Ricardo, Jr	1270 Montgomery Ave	05447	1953	
2	Tim Cratchit	505 Charles St	05403	2000	> 99.9
	Bob Cratchit	505 Charles St	05403	1975	
3	Ellen Ripley	1629 Clifford Farm Rd	05065	1960	92.0
	Sarah Connor	1629 Clifford Farm Rd	05065	1961	
4	Elizabeth Lemon	160 Riverside Dr	05443	1974	99.2
	Elizabeth Bennett	160 Riverside Dr	05443	1941	
5	Mary Poppins	1115 Creek Rd Apt 1	05032	1942	99.4
	Mary Wollstonecraft	1115 Creek Rd	05032	1941	
6	Michelle Yeoh	521 Warren Rd Unit B	05871	1962	> 99.9
	Michelle Obama	521 Warren Rd Unit A	05871	1964	
7	Nathan Chen	208 Pearl Ave	05478	1992	89.5
	Richard Cheney	46 Pearl Ave	05478	1991	
8	Martin Brody	200 Sharkyville Rd	05676	1937	> 99.9
	Ellen Brody	200 Sharkyville Rd	05676	1942	

Apartment buildings lead to matching difficulties when the unit number is omitted from a survey participant’s address or the voter file, which is not particularly unusual. If the sampled person shares a first name with another resident of the building, and the unit number is missing from one record, the match probability can be > 99% (pair 5, matching Mary Poppins to Mary Wollstonecraft). Even if different unit numbers are included, the match score can be high for people with different names and ages (pair 6 showing > 99.9% for Michelle Yeoh and Michelle Obama). And people who live in different houses on the same street can match when their names barely resemble each other (pair 7, where Nathan Chen is an 89.5% match to Richard Cheney at a different address).

Gender errors prevent correct matching when comparisons are blocked on gender, following Enamorado et al. method (2018). Identical records except for opposite genders will never match, and if one record has a gender error, it can generate a high-confidence false match (pair 8, where Martin Brody matches > 99.9% to his wife Ellen despite different names and ages).

These examples of invalid matching outcomes illustrate types of errors observed in probabilistic matching results using the ANES dataset before cleri-

cal review. What is the prevalence of such troublesome matches? In Enamorado and Imai's (2018) review of the 2016 ANES data, 100% of cases with match probabilities below 85% were judged incorrect (false positives), and 10% of those judged incorrect had probabilities exceeding 99.9%. In the 2020 ANES data, ANES staff systematically reviewed all the matched pairs, comparing each ANES respondent to the fastLink-identified best match to a commercial voter file record (ANES, 2022). As elaborated in Online Appendix 3, this review found that a majority (55.3%) of cases with match probabilities in the range of approximately 85% to 99.9999% (about 19% of the sample) were incorrect and that most (79.7%) matches with probabilities less than 80.0% (about 16% of the sample) were incorrect. Even matches with 99.85% to 99.95% probabilities were correct only 53.6% of the time. Both reviews show that fastLink's results need to be further processed to remove false positives. This can be accomplished through clerical review.

Of course, all matching procedures make errors. I am able to highlight these only because fastLink's authors made the software freely available. Proprietary matching systems are more difficult to assess directly, but, as noted above, the Pew Research Center compared them and found varying results across five firms that point to substantial error rates and unreliability (Igielnik et al., 2018).

Clerical Review

The matching errors detailed above show the need to review all of fastLink's matches. Enamorado and Imai (2019) recommended doing so selectively—they favored it for ANES but not CES because CES VV estimated a turnout rate close to the official one without clerical review (and, practically, the CES dataset was perhaps too large to review manually). But as we have seen, the potential for offsetting errors means that correspondence between the validated turnout rate and the actual turnout rate does not demonstrate VV accuracy. Objective clerical review of fastLink results is generally warranted. The procedure should be reproducible and have its reliability reported. I propose an objective review protocol, summarized in Appendix B and detailed in Online Appendix 3, to tally the differences of minor significance, such as a one-character typo or using a recognizable nickname, and of major significance, such as a different address or last name. A simple formula applied to the tallies classifies each pair as correctly or incorrectly matched. Results from two coders working independently to apply the objective criteria to the ANES 2020 sample were reliable (98.3% agreement, Cohen's $\kappa = .954$), and both coders' objective scores also corresponded very closely to collective subjective judgments by a group of 3 coders who examined the matches (97.6% and 98.1% agreement, Cohen's $\kappa = .934$ and $.948$, respectively, for reliability between the subjective judgment and the objective codes of the two coders). This demonstrates that objective coding methods are reliable (or

consistent), because of the high agreement and Cohen's κ statistics, and that they appear valid (or true) because they correspond to the judgments of humans looking closely at the cases independently. Matches suggested by probabilistic software should be reviewed in this way.

Match Probability Scores Are Not Calibrated Probabilities Or Valid Weights

Match probability is calculated by fastLink using Bayes' Rule and a maximum likelihood estimator. As I have described, these probabilities can be problematic. The scores are not the probability that the match is correct; to confirm a hypothesis that the scores are real probabilities would require a study showing that, for instance, among the cases with match scores of p , approximately that proportion were correct matches. We have two independent rejections of the hypothesis. In Enamorado and Imai's (2019) manual review of ANES 2016 matches, 100% of cases with match probabilities below 85% were deemed (subjectively) incorrect, so for each probability below 85%, instead of finding the given percentage valid, the percentage found valid was 0. The ANES review of its 2020 data, described above, had the same outcome for the probabilities. Moreover, they cannot be actual match probabilities because they are not constrained to sum to 1; for most cases, fastLink produces multiple possible matches, and the sums of the match probabilities for each case are often > 1 . It is not unusual to have two matches $> .99$.

Enamorado et al. (2018, 2019) argue for weighting based on the match probabilities. The argument for such weighting is that even if the probabilities are poorly calibrated, results may be valid if the probabilities are unbiased. Yet, given that the match scores do not describe the probability that the match is correct, they are not appropriate weights for data analysis. Classical sampling weights equal (or are proportional to) the inverse of the selection probability, where selection probability is the actual probability of including the case in the sample. Weighting on the match probabilities reflects an analogous assumption that they reflect the actual probability that a respondent with a given score was a voter; if several cases have $p = .5$, and 50% of these actually voted, then the weight yields an unbiased estimate of turnout. But as we have seen, this is not what the match probabilities mean. In practice, weighting on the probability introduces error to the estimates, while reliable clerical review results have better face validity. A future study could quantify the error in these probabilities and propose calibration adjustments.

Many “Not-Founds” Appear to Be Voters

The epigraph of this paper asked what we should make of voters who are “not found.” This is an important factor in VV’s accuracy at the individual level. The question cannot be answered definitively, but cross-checking independent VVs provides evidence. The ANES 2020 study did so. Probabilistic VV for the ANES sample was performed by NORC using a commercial voter database and was separately performed by ANES staff using a different database with clerical review as described above (ANES, 2022). The two methods had different turnout results for 18% of ANES respondents, including 4% where one method found no match and the other found a voter. Of the cases not found by one vendor ($n = 618$), 50% were linked to a voter by the other vendor. Of those not found by NORC but found by the ANES clerical review ($n = 72$), a majority (69%) were linked to a voter. This last procedure should have few false positives and shows that probabilistic matching does not allay decades-old concerns about not-found being voters.

An Interpretive Framework to Estimate Error Sizes

The live controversy in the VV literature is between the claim that VV is largely accurate (e.g., Ansolabehere & Hersh, 2012; Enamorado & Imai, 2019) and the claim that the appearance of accuracy is an illusion caused by matching failures (Berent et al., 2016). The data errors and matching problems already described show that matching failures are a serious concern, but we can go further to assess their effects on VV’s credibility. In this section I consider in a new way how three kinds of factors jointly constrain each other to affect VV’s accuracy. The aim, given what we know about matching errors and data errors, is to say what we can deduce about the accuracy of vote validation and how sensitive VV’s accuracy is to assumptions that are contested in the literature. This section proceeds as follows. First, I operationally define turnout accuracy indicators. Next, I detail how, theoretically, three complex kinds of factors—assumptions about matching accuracy, assumptions about sample proportions in an error matrix, and observed outcomes—create joint constraints that provide insight about VV’s accuracy. Then I justify a range of assumptions in this complex of factors. The last step is to input credible values into the system of equations and calculate feasible outcomes for the accuracy of VV compared to self-reported turnout.

By analogy, given two side lengths and one angle of a triangle, trigonometry reveals the other length and angles. If we can only approximate the “givens,” we cannot calculate the other properties precisely, but we can approximate them, too. Similarly, although many properties of VV are not precisely known, the evidence establishes ranges for certain factors that constrain VV’s accuracy to fall within a calculable range. This section elaborates and applies this idea.

Turnout Accuracy Indicators

Previous studies have used the validated turnout rate's correspondence with the true turnout rate as a quality indicator, but as Berent et al. (2016) pointed out, this correspondence is no indicator of success unless we assume what needs to be proven, that matching gives valid results. I report four turnout accuracy indicators: proportion correct, correlation, sensitivity, and specificity. The proportion correct is simply the number of correct turnout data points divided by the total number (correct + incorrect). While it has the virtue of simplicity, as a single metric summarizing a two-dimensional result (eliding true positives and true negatives), it can be misleading. It is also prevalence dependent: if 80% of a sample voted, then an estimate that says everyone voted is 80% correct but not informative. Correlation, in this case ϕ (ϕ), is the association between the truth and estimated or validated turnout. It has the virtues of not being prevalence-dependent and of accounting for both the positive and negative elements of the error matrix (defined below), which make it one of the preferable summary statistics (Chicco, 2021). Sensitivity and specificity differentiate the two dimensions in accuracy and correlation. Sensitivity is the true positive rate, or probability of detection: the probability of finding a vote when the respondent is a true voter. Specificity is the true negative rate, also known as selectivity: the probability of finding no vote when the respondent is a true non-voter. The ideal test is both sensitive and specific, but for some purposes one may be more important than the other. When testing patients for deadly diseases, the sensitivity is usually more important, because the consequences of a false negative may be worse than a false alarm. Specificity, on the other hand, is better for excluding cases that do not have the condition of interest, such as separating the innocent from the guilty before punishment.

Joint Constraints—Assumptions about Sample Proportions in the Error Matrix

The literature has usually treated VV accuracy as if it were one-dimensional, consisting of the difference between the estimated turnout rate and the true rate. But, like any dichotomous classification problem, it is really two-dimensional, because we can separately consider accuracy for voters and for non-voters. An error matrix is a 2x2 contingency table that describes true or false outcomes from a classification attempt. For our purposes the error matrix breaks down true turnout status by self-reported turnout status to give four proportions: (a) respondents who are true voters who self-report voting, who may be termed "true positives" (TP) for turnout reports; (b) true voters who falsely report not voting (false negatives; FN); (c) true non-voters who falsely report voting (false positives; FP); and (d) true non-voters who report not voting (true negatives; TN).

These proportions are unknown and may vary considerably depending on the sample, context, and question wording. They provide critical analytical leverage when considered in the context of their joint constraints with matching accuracy and observed outcomes.

Joint Constraints—Assumptions about Matching Accuracy

Attempted record linkage has three possible outcomes: a correct match, a non-match, or a false match. The probability of each may differ depending on whether the survey respondent is or is not a true voter, given that true voters are registered and true non-voters are often unregistered, and that true voters and non-voters differ in demographic and other characteristics that are correlated with the quality of voter file data. We should therefore consider six conditional probabilities: the probability of a correct match, a non-match, or a false match, each respectively conditional on being a true voter or true non-voter. Additionally, we must consider that a false match may point to a voter or a non-voter. This makes eight conditional probabilities for matching, which are inter-related in straightforward ways. For example, the probability of a correct match conditional on being a true voter is equal to 1 minus the probabilities of being a non-match or false match (also among true voters). These relationships are detailed in Table A1 of Appendix A.

Joint Constraints—Observed Outcomes

Several observed characteristics of the data jointly constrain the possible range of the assumptions about matching accuracy and the error matrix. These include the self-reported turnout rate, the match rate (i.e., the proportion of the sample successfully matched to the voter file), the validated turnout rate (i.e., the proportion of the sample that the voter file indicates voted, with unmatched cases included in the denominator), the probability of a self-reported vote conditional on a validated vote, the probability of a self-reported vote conditional on not validating turnout, and the probability of a validated vote conditional on a reported vote. These proportions are usually known for any study on which VV has been performed (see Table 3), and we can use them to explore the boundaries they collectively impose on matching accuracy. For example, the self-reported turnout rate is:

$$\text{Self-reported turnout} = TP + FP.$$

Using the notation $\Pr(A \mid B)$ to mean the probability of A conditional on B, the match rate is:

$$\begin{aligned} \text{Match rate} = & [\Pr(\text{Correct match} \mid \text{True voter}) + \Pr(\text{False match} \mid \text{True voter})] \times (TP + FN) \\ & + [\Pr(\text{Correct match} \mid \text{True non-voter}) + \Pr(\text{False match} \mid \text{True non-voter})] \times (FP + TN) \end{aligned}$$

The validated turnout rate is:

$$\begin{aligned} \text{Validated turnout rate} = & TP \times [\Pr(\text{Correct match} \mid \text{True voter})] \\ & + TP \times [\Pr(\text{False match} \mid \text{True voter}) \times \Pr(\text{Finding non-voter} \mid \text{False match})] + FN \times [\Pr(\text{Correct match} \mid \text{True voter})] \\ & + FN \times [\Pr(\text{False match} \mid \text{True non-voter}) \times \Pr(\text{Finding voter} \mid \text{False match})] \\ & + FP \times [\Pr(\text{False match} \mid \text{True non-voter}) \times \Pr(\text{Finding voter} \mid \text{False match})] \\ & + TN \times [\Pr(\text{False match} \mid \text{True non-voter}) \times \Pr(\text{Finding voter} \mid \text{False match})] \end{aligned}$$

Table 3

Observed Vote Validation Outcomes From Three Studies

Outcome	CCES 2008	BKL	ANES 2016
Self-reported turnout rate	84.3	87.6	75.1
Match rate	83.7	65.1	68.0
Validated turnout rate	68.5	59.3	59.8
$\Pr(\text{Reported vote} \mid \text{Validated vote})$	99.1	100.0	98.4
$\Pr(\text{Reported vote} \mid \text{Not validated vote})$	52.0	–	41.2
$\Pr(\text{Validated vote} \mid \text{Reported vote})$	80.6	–	77.6

Note. CCES estimates from Ansolabehere & Hersh, 2012. See Online Appendix 1 for ANES estimates. “BKL” estimates from Berent, Krosnick, and Lupia 2016.

– not reported.

These and the other observed outcomes jointly constrain each other, the matching accuracy, and the error matrix, revealing ranges of possible results given specified assumptions. For example, it is evident from the match rate formula above that our priors for TP, TN, FP, FN, and the included conditional probabili-

ties will constrain the match rate to the extent that these priors are limited in their own ranges. Formulae for the other outcomes in Table 3 are shown in Table A1 of Appendix A.

Method: Credible Value Inputs

This analysis framework operates by defining observed or credible values for the “joint constraints” described above and then using these values to calculate turnout accuracy indicators. Credible values are based on the evidence about errors in government records, commercial voter files, and survey datasets that was reviewed earlier. Establishing boundaries for the credible values is challenging and unavoidably subjective due to limitations in the empirical evidence. My approach is to define optimistic and pessimistic scenarios and use these as boundaries for best-case and worst-case outcomes of VV. Optimistically, we could assume that the effects of errors on matching accuracy are small and that errors are compounded in ways favorable to the overall match rate, such as assuming that the cases with a gender reversal (where a match is impossible when fastLink blocks on gender) are often also the cases where the panel respondent substitution occurs (which also makes a correct match effectively impossible). Such favorable assumptions support optimistic rates for match failures. More pessimistically, we may assume that compounding errors are unfavorable to the overall match rate (e.g., gender reversal and respondent substitution sometimes occur in different sets of respondents) and that errors cause relatively large reductions in match rates.

In Table 4 I suggest credible pessimistic and optimistic values for conditional probabilities and the error matrix. Online Appendix 1 discusses and justifies the selection of credible values in detail. When giving results, I will select scenarios consisting of a set of values relatively favorable for VV and a set of values relatively favorable for self-reports, which approximate the results from Table 3, and which as much as possible fall within the optimistic and pessimistic ranges defined by Table 4, and then I will consider the turnout data accuracy indicators for those outcomes. In the interest of brevity and applicability to the literature, I examine two scenarios each tailored to the results from the VV conducted for the 2008 CCES (Ansolabehere & Hersh, 2012), from Berent et al.’s VV study (2016), and from the 2016 ANES (Enamorado & Imai, 2019).

Method: Calculation

To estimate the credibility of VV, I apply credible values to the parameters described above and use these inputs to calculate the accuracy of VV. The assumptions, outcomes, and quality indicators are given in Table 5 (and their formulae are given in Table A1 of Appendix A). The challenging aspect of this

is to select values for these 12 parameters that are allowable, given their complex mutual constraints, and consistent with the empirical results in Table 3. I programmed a simulation that randomly assigns values to the matching probabilities and the error matrix. (Code is in Online Appendix 1.) The program generates scenarios for these assumptions, calculates the observed outcomes, and selects only those that are close to the empirical results (within 1 percentage point or less on all six outcomes). For example, to correspond to the CCES 2008 results, the dozen inputs must collectively give results within 1 percentage point of a reported turnout rate of 84.3%, a match rate of 83.7%, a validated turnout of 68.5%, 99.1% of respondents reporting voting conditional on validation, 52.0% reporting voting conditional on not-validated turnout, and 80.6% with validated turnout conditional on self-reported turnout, thus approximating the known outcomes from the validation of the 2008 CCES by Ansolabehere and Hersh (2012). This exercise does not examine all possible scenarios that could lead to the observed outcomes, because the number is practically infinite, but it illustrates feasible examples consistent with known data in an underdetermined system.

Table 4

Credible Ranges for Vote Validation's Conditional Probabilities and Error Matrix

Variable	Pessimistic value	Optimistic value
Assumptions about matching (conditional probabilities)		
Pr(Correct match True voter)	.76	.91
Pr(Non-match True voter)	.15	.03
Pr(False match True voter)	.21	.00
Pr(Correct match True non-voter)	.20	.70
Pr(Non-match True non-voter)	.77	.04
Pr(False match True non-voter)	.41	.03
Pr(Finding voter False match)	.54	.94
Pr(Finding non-voter False match)	.46	.06
Assumptions about sample proportions (error matrix)		
True voter self-reports voting (TP)	.58	.89
True voter reports not voting (FN)	.02	.00
True non-voter reports voting (FP)	.17	.06
True non-voter reports not voting (TN)	.24	.05

Note. Optimistic and Pessimistic values are supported in Online Appendix 1.

Results

Table 5 gives results for two scenarios each conforming to the outcomes of the CCES 2008, the Berent et al. (2016) study (referred to as “BKL”), and the ANES 2016.

Table 5

Turnout Accuracy of Self-Reports and Validation Under Six Sets of Assumptions

Variable	CCES 2008		BKL		ANES 2016	
	1	2	3	4	5	6
Assumptions about matching (conditional probabilities)						
Pr(Correct match True voter)	.905	.839	.682	.684	.856	.809
Pr(Non-match True voter)	.055	.096	.318	.314	.102	.149
Pr(False match True voter)	.040	.065	.001	.002	.042	.042
Pr(Correct match True non-voter)	.502	.551	.496	.496	.227	.270
Pr(Non-match True non-voter)	.433	.364	.464	.494	.735	.699
Pr(False match True non-voter)	.065	.085	.040	.010	.038	.030
Pr(Finding voter False match)	.586	.584	.756	.590	.668	.610
Pr(Finding non-voter False match)	.414	.416	.244	.410	.332	.390
Assumptions about sample proportions (error matrix)						
True voter self-reports voting (TP)	.721	.774	.844	.839	.661	.704
True voter reports not voting (FN)	.002	.001	.009	.013	.005	.003
True non-voter reports voting (FP)	.119	.064	.038	.036	.089	.049
True non-voter reports not voting (TN)	.158	.161	.109	.112	.244	.244
Observed outcomes						
Reported turnout rate	.840	.838	.882	.875	.750	.753
Match rate	.840	.844	.661	.659	.686	.690
Validated turnout rate	.677	.682	.587	.584	.588	.589
Pr(Reported vote Validated vote)	.988	.987	.983	.984	.982	.988
Pr(Reported vote Not validated vote)	.529	.518	.738	.722	.420	.416
Pr(Validated vote Reported vote)	.806	.814	.665	.668	.784	.782
Turnout accuracy indicators						
Self-reports						
Proportion correct	.879	.935	.953	.951	.905	.948
Sensitivity: Pr(Reported vote True voter)	.997	.999	.989	.985	.992	.996

Table 5 *continued*

Specificity: Pr(No reported vote True non-voter)	.570	.716	.741	.757	.733	.833
Correlation (ϕ)	.693	.810	.802	.796	.790	.874
<i>Validation</i>						
Proportion correct	.933	.885	.725	.731	.904	.871
Sensitivity: Pr(Validated vote True voter)	.928	.877	.682	.685	.884	.835
Specificity: Pr(No validated vote True non-voter)	.962	.950	.970	.994	.975	.982
Correlation (ϕ)	.855	.748	.469	.489	.826	.757

The CCES results in column 1 assume that validation correctly matches true voters 91% of the time ($\text{Pr}(\text{Correct match} \mid \text{True voter}) = .905$). In tandem with the observed match rate of 84% (match rate = .840), combined with an assumption that a majority of the sample consists of voters, this constrains the match rate on true non-voters to be much lower; it is, of course, impossible for a sample with 72% true turnout ($\text{TP} = .721$) to have an overall match rate of 84% and a match rate for true voters of 91% unless the match rate on true non-voters is much lower. In Column 1 the match rate for non-voters is 50% ($\text{Pr}(\text{Correct match} \mid \text{True non-voter}) = .502$). The scenario also assumes that 0.2% voted but reported they did not vote ($\text{FN} = .002$), 11.9% over-reported ($\text{FP} = .119$; a high figure relative to many estimates, but one that should work to VV's favor), and 15.8% truthfully reported not voting ($\text{TN} = .158$).

With these assumptions in place, we find that self-reported turnout is, overall, 87.9% correct (.879 in Table 5), with excellent sensitivity of 99.7% and poor specificity of 57.0%. By comparison, the validated results are 93.3% correct, with sensitivity of 92.8% and much better specificity of 96.2%. The correlation is .693 for self-reported turnout and .855 for validated turnout. This shows that, unless sensitivity is a highly weighted priority, validation can be more accurate overall than self-reports and that CCES 2008 validation may have been. Note, again, this is only one of many possible scenarios consistent with published results.

Column 2 illustrates a contrasting scenario, more favorable to self-reports and with higher validation error rates. It assumes that matching is 83.9% accurate for true voters, 55.1% accurate for true non-voters, and, compared to Column 1, it assumes more true voters in the sample (77.4%) and fewer over-reporters (6.4%). These assumptions, like all scenarios shown in the table, closely fit the study's observed results. In this case, self-reported data have better overall correctness than validated votes (93.5% compared to 88.5%), better sensitivity (99.9% compared to 87.7%), and a higher correlation (.810 compared to .748), but lower specificity (71.6% compared to 95.0%).

Columns 3 through 6 collectively tell the same story as the first two, with columns 3 and 4 (BKL) using data from Berent et al.'s (2016) study and columns 5 and 6 using the ANES 2016 data validated by Enamorado and Imai (2019). For ANES and CCES there are scenarios, consistent with published data and credible assumptions, that produce either validation outperforming self-reports or self-reports outperforming validation. Thus, the data from these studies do not establish which is better, and the conclusion depends on which assumptions (within the feasible set) one adopts.

The BKL example differs somewhat from the other two. In the CCES and ANES cases, the assumed conditional probabilities are all within the bounds established in Table 4, but for the BKL case, I could find no inputs that were within all these bounds (though I tried tens of millions) and had to relax some of them.² Also, consistent with the main finding of BKL's paper, I could find no example where VV was more accurate than self-reports, except for specificity. That is, in the study where the authors found VV was not accurate, the results are consistent with that finding. More interestingly, in all three studies, self-reports had better sensitivity (counting voters as voters) and validation had better specificity (excluding non-voters).

Discussion and Conclusions

Validating self-reported turnout is desirable because over-reporting is common. Yet I have reviewed several reasons to treat the results of probabilistic vote validation with caution. Errors in voter files and survey datasets increase the chances of two adverse outcomes: false negatives and false matches. False negatives are misinterpreted as non-voting, while false matches, in expectation, bias results toward the mean in the voter file. Together, false negatives and false matches depress validated turnout estimates, which, as argued by Berent et al. (2016), probably creates an illusion of accuracy in VV results. Moreover, the matching procedures are themselves error-prone and give match scores that quantify linkage confidence but should not be treated as calibrated probabilities or weights. Clerical review of these results is needed and should follow a demonstrably reliable and reproducible procedure such as the one summarized in Appendix B and detailed in Online Appendix 3.

Commercial voter files and open-source matching software have made VV cheaper, more accessible, and more reproducible, but it is not clear how much accuracy has improved. Using credible assumptions for matching accuracy and reporting error, self-reported turnout has higher sensitivity (it better identifies

² It is unclear why this occurred. It is possible that one or more true values for the BKL study is outside the "credible range" I defined for inputs (Table 4).

true voters), while validated turnout has higher specificity (it better excludes non-voters). Which has better overall accuracy depends on unknowns. Thus, this review indicates that a uniform preference for current VV techniques over self-reports is inappropriate. Instead, I offer three suggestions for data collectors, followed by three for analysts.

Survey researchers who perform VV need a method for reliable clerical review of linkage results. Coding standards that can be followed by research staff and that have demonstrably high inter-coder reliability are a start (see Appendix B). Better yet, a computer script to implement clerical review would make the results not only reproducible but more cost-effective with large samples.

Since survey and voter data files have errors that impede record linkage, researchers need better evidence of fastLink's accuracy and bias when using data riddled by realistic errors. Such a test could be performed by sampling from a voter database, randomly introducing errors to the sample to simulate a realistic assortment of errors of misreporting, data entry, and item nonresponse, and then matching the perturbed sample back to the database to check the effects of the data errors on the matching, including rates of false positives and false negatives, the effects on the validated turnout rate, and the introduction of bias to the results.

Online validation during interviews could improve the process. The standard method is to match survey data to voter files after data collection is done. In the age of online surveys with address-based sampling or established panels, this approach should be obsolescent. Instead, by pre-matching sampled addresses to voter databases, all potential matches associated with the sampled address can be pre-loaded into a survey system. Once the respondent confirms their address and provides a name, gender, and date of birth, the system can select within the short list of pre-loaded potential matches for individual confirmation of the best one, with a message such as, "Public records show this person living at the address where we contacted you. Is this you, is this someone else you know, or is this someone you don't know at all?" Respondents might also be shown a list of potential matches and be asked which record is theirs. This is now technically feasible and has the potential to improve matching results by validating record linkages.

Data analysts should not weight probabilistic VV data by match probability scores, because the scores are not valid probability estimates. Valid probability estimates potentially could be estimated by calibrating the posterior probabilities, so they sum to 1 for each case and incorporate linkage uncertainty from Bayesian priors.

Analysts choosing between self-reported or validated turnout face tradeoffs and should be deliberate about their priorities. VV errors most often misclassify real voters as non-voters, while self-reporting errors most often misclassify real non-voters as voters. Both errors likely attenuate the observed differences

between voters and non-voters, making predictors of turnout appear weaker than they really are. But how they do this probably differs because the different sources of error affect different turnout predictors. Validation errors are often associated with income, home tenure, residential mobility, name changes, political disengagement, and incomplete or inaccurate identifying information (perhaps from those low in social trust or vain about their age) (Igielnik et al., 2018; Jackman & Spahn, 2021). Self-reporting errors, driven especially by social desirability, have been associated with income, education, partisan strength, political engagement, and church attendance (Ansolabehere & Hersh, 2012; Bernstein et al., 2001). Thus, both validation error and reporting error cause bias in predictors of turnout and representation of voters, but different biases. Since self-reporting is very sensitive but less specific, it is biased toward overestimating turnout, overestimating nonresponse bias, and distorting the estimated electorate on the variables associated with over-reporting. On the other hand, vote validation is specific but less sensitive, and bias is toward underestimating turnout in the sample, underestimating nonresponse bias, and distorting the electorate on the characteristics associated with validation error. When both measures are available, it is worth considering which is more accurate for predictors of interest and comparing results as a robustness check.

Finally, scholars should maintain their skepticism. Commercial voter files and recent work such as Enamorado and Imai's (2019) have moved the field toward better validation by making it more accessible, transparent, and reproducible, but VV has not solved the problem of overreporting. It has added useful data that still have serious limitations. Reviewing sources of error and inconsistency in VV three decades ago, Traugott et al. (1992, p. 12) concluded that "the results of voter validation efforts should be treated with some care." Just so.

Acknowledgments

The author thanks Ted Enamorado for feedback on a draft of the paper and the Vermont Secretary of State's office for support with the Vermont voter file.

Data and Code Availability

Code to reproduce the analyses is included in the Online Appendices. Data to reproduce ANES analyses are available from the American National Election Studies (https://electionstudies.org/anes_timeseries_2020_spss_VoterValidation/). Data to reproduce the analysis of Vermont voter data are available from the Vermont Secretary of State.

Conflict of Interest

The author declares no conflict of interest.

Ethics

The ANES data collection and processing were approved by the Institutional Review Boards of the University of Michigan and Stanford University.

Funding

This work was funded by the National Science Foundation under grant numbers SES-1835721 and SES-2209438 to the University of Michigan and grant SES-1835022 to Stanford University.

References

- Achen, C., & Blais, A. (2016). Intention to vote, reported vote, and validated vote. In J. A. Elkind & D. M. Farrell (Eds.), *The Act of Voting*. Routledge.
- ANES. (2022). *User's guide and codebook for the ANES 2020 Time Series Voter Validation Supplemental Data* [Computer file]. The University of Michigan and Stanford University. https://electionstudies.org/wp-content/uploads/2022/10/anes_timeseries_2020_userguidecodebook_VoterValidation.pdf
- Ansolabehere, S., & Hersh, E. (2010). *The quality of voter registration records: A state-by-state analysis*. Harvard University. https://vote.caltech.edu/documents/187/reg_quality_report_8-5-10.pdf
- Ansolabehere, S., & Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, 20(4), 437-459. <https://doi.org/10.1093/pan/mps023>
- Berent, M. K., Krosnick, J. A., & Lupia, A. (2016). Measuring voter registration and turnout in surveys: Do official government records yield more accurate assessments? *Public Opinion Quarterly*, 80(3), 597-621. <https://doi.org/10.1093/poq/nfw021>
- Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly*, 65(1), 22-44. <http://www.jstor.org/stable/3078784>
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-Million person experiment in social influence and mobilization. *Nature*, 489, 295-298. <https://doi.org/10.1038/nature11421>
- Chicco, D., Starovitev, V., & Jurman, G. (2021). The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access*, 9, 47112-47124. <https://doi.org/10.1109/ACCESS.2021.3068614>

- Dahlggaard, J. O., Hansen, J. H., Hansen, K. M., & Bhatti, Y. (2019). Bias in self-reported voting and how it distorts turnout models: Disentangling nonresponse bias and overreporting among Danish voters. *Political Analysis*, 27(4), 590-598. <https://doi.org/10.1017/pan.2019.9>
- DeBell, M., Amsbary, M., Brader, T., Brock, S., Good, C., Kamens, J., Maisel, N., & Pinto, S. (2022). *Methodology report for the ANES 2020 Time Series Study*. https://electionstudies.org/wp-content/uploads/2022/08/anes_timeseries_2020_methodology_report.pdf
- DeBell, M., Krosnick, J. A., Gera, K., Yeager, D. S., & McDonald, M. P. (2020). The turnout gap in surveys: Explanations and solutions. *Sociological Methods and Research*, 49(4), 1133-1162. <https://doi.org/10.1177/0049124118769085>
- DeBell, M., Maisel, N., Brader, T., & Meldener, V. (2018). Nonresponse bias in a nationwide dual-mode survey. Presented at the International Total Survey Error Workshop 2018, Durham, NC. https://dism.duke.edu/files/2020/05/debell_maisel_brader_meldener.pptx
- DeBell, M., Maisel, N., & Brader, T. (2022). *Nonresponse biases in the American National Election Studies* [Unpublished manuscript].
- Enamorado, T., & Imai, K. (2019). Validating self-reported turnout by linking public opinion surveys with administrative records. *Public Opinion Quarterly*, 83(4), 723-748. <https://doi.org/10.1093/poq/nfz051>
- Enamorado, T., Fifield, B., & Imai, K. (2018). User's guide and codebook for the ANES 2016 Time Series Voter Validation Supplemental Data [Computer file]. https://electionstudies.org/wp-content/uploads/2018/03/anes_timeseries_2016voteval_userguidecodebook.pdf
- Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2), 353-371. <https://doi.org/10.1017/S0003055418000783>
- Fowler, J. H., Baker, L. A., & Dawes, C. T. (2008). Genetic variation in political participation. *American Political Science Review*, 102(2), 233-248. <https://doi.org/10.1017/S0003055408080209>
- Fraga, B. L. (2016). Candidates or districts? Reevaluating the role of race in voter turnout. *American Journal of Political Science*, 60(1), 97-122. <https://doi.org/10.1111/ajps.12172>
- Igielnik, R., Keeter, S., Kennedy, C., & Spahn, B. (2018). Commercial voter files and the study of U.S. politics. Pew Research Center. <https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2018/12/FINAL-Voter-File-Report-2.15.18.pdf>
- Jackman, S., & Spahn, B. (2019). Why does the American National Election Study overestimate voter turnout? *Political Analysis*, 27(2), 193-207. <https://doi.org/10.1017/pan.2018.36>
- Jackman, S., & Spahn, B. (2021). Politically invisible in America. *PS: Political Science and Politics*, 54(4), 623-629. <http://doi.org/10.1017/S1049096521000639>
- Miller, J. D., Kalmbach, J., Woods, L. T., & Cepuran, C. (2021). The accuracy and value of voter validation in national surveys: Insights from longitudinal and cross-sectional studies. *Political Research Quarterly*, 74(2), 332-347. <https://doi.org/10.1177/1065912920903432>
- Nyhan, B., Skovron, C., & Titunik, R. (2017). Differential registration bias in voter file data: A sensitivity analysis approach. *American Journal of Political Science*, 61, 744-760. <https://doi.org/10.1111/ajps.12288>
- Parry, H. J., & Crossley, H. M. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, 14(1), 61-80. <https://doi.org/10.1086/266150>

- Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105(492), 1418-1430. <https://doi.org/10.1198/jasa.2010.ap09485>
- Pew Center on the States. (2012). Inaccurate, costly, and inefficient: Evidence that America's voter registration system needs an upgrade. Issue Brief. https://www.pew-trusts.org/-/media/legacy/uploadedfiles/pes_assets/2012/pewupgradingvoterregistrationpdf.pdf
- Presser, S., Traugott, M. W., & Traugott, S. (1990). Vote 'over' reporting in surveys: The records or the respondents? Paper presented at the International Conference on Measurement Errors, Tucson, AZ. <https://electionstudies.org/wp-content/uploads/2018/04/nes010157.pdf>
- Rosenstone, S. J., & Legee, D.C. (1994). An update on the National Election Studies. *PS: Political Science and Politics* 27, 693-698. <https://doi.org/10.1017/S1049096500048307>
- Selb, P., & Munzert, S. (2013). Voter overrepresentation, vote misreporting, and turnout bias in postelection surveys. *Electoral Studies*, 32(1), 186-196. <https://doi.org/10.1016/j.electstud.2012.11.004>
- Tourangeau, R., Groves, R. M., & Redline, C. D. (2010). Sensitive topics and reluctant respondents: Demonstrating a link between nonresponse bias and measurement error. *Public Opinion Quarterly*, 74(3), 413-432. <https://doi.org/10.1093/poq/nfq004>
- Traugott, S. (1989). Validating self-reported vote: 1964-1988 [Paper presentation]. Annual meeting of the American Statistical Association, Washington, DC. <https://electionstudies.org/wp-content/uploads/2018/04/nes010152.pdf>
- Traugott, M. W., Traugott, S., & Presser, S. (1992). Revalidation of self-reported vote. Presented at annual meeting of the American Association for Public Opinion Research, St. Petersburg Beach, FL. <https://electionstudies.org/wp-content/uploads/2018/04/nes010160.pdf>

Appendix A

Formulas

Table A1

Formulas for Accuracy of Matching Outcomes in Table 5

Variable		Formula
Assumptions about matching (conditional probabilities)		
<i>a</i>	Pr(Correct match True voter)	Given
<i>b</i>	Pr(Non-match True voter)	Given
<i>c</i>	Pr(False match True voter)	$1 - a - b$
<i>d</i>	Pr(Correct match True non-voter)	Given
<i>e</i>	Pr(Non-match True non-voter)	Given
<i>f</i>	Pr(False match True non-voter)	$1 - d - e$
<i>g</i>	Pr(Finding voter False match)	Given
<i>h</i>	Pr(Finding non-voter False match)	$1 - g$
Assumptions about sample proportions (error matrix)		
TP	True voter self-reports voting (True Positive)	Given
FN	True voter reports not voting (False Negative)	Given
FP	True non-voter reports voting (False Positive)	Given
TN	True non-voter reports not voting (True Negative)	$1 - (TP+FN+FP)$
Observed outcomes		
	Reported turnout rate	$TP + FP$
	Match rate	$(a+c)(TP+FN) + (d+f)(FP+TN)$
<i>i</i>	Validated turnout rate	$TP \times a + TP(c \times h) + FN(a) + FN(f \times g) + FP(f \times g) + TN(f \times g)$
<i>j</i>	Reported voter	$TP \times a + TP(c \times h) + FP(f \times g)$
	Reported not voting	$FN \times a + FN(f \times g) + TN(f \times g)$
<i>k</i>	Validated non-voter	$TP(c \times g) + TP \times b + FN(f \times h) + FN \times b + FP \times d + FP(f \times h) + FP \times e + TN \times d + TN(f \times h) + TN \times e$
<i>m</i>	Reported voter	$TP(c \times g) + TP \times b + FP \times d + FP(f \times h) + FP \times e$
	Reported non-voter	$FN(f \times h) + FN \times b + TN \times d + TN(f \times h) + TN \times e$
<i>p</i>	Pr(Reported vote Validated vote)	j / i
<i>q</i>	Pr(Reported vote Not validated vote)	m / k
<i>r</i>	Pr(Validated vote Reported vote)	$i / (j+m)$

Table A1 *continued*

Variable		Formula
Turnout accuracy indicators		
Self-reported turnout		
	Proportion correct	TP+TN
	Sensitivity: Pr(Reported vote True voter)	TP / (TP+FN)
	Specificity: Pr(No reported vote True non-voter)	TN / (FP+TN)
	Correlation (ϕ)	$(TP \times TN - FP \times FN) / [(TP+FP)(TP+FN)(FP+TN)(FN+TN)]^{.5}$
Validated turnout		
	Proportion correct	$TP \times a + TP \times c \times h + FN \times a + FN \times f \times g + FP \times d + FP \times f \times h + FP \times e + TN \times d + TN \times f \times h + TN \times e$
TPR	Sensitivity: Pr(Validated vote True voter)	$TP_v / (TP_v + FN_v)$
TNR	Specificity: Pr(No validated vote True non-voter)	$(FP \times d + FP \times f \times h + FP \times e + TN \times d + TN \times f \times h + TN \times e) / (FP \times d + FP \times f \times g + FP \times f \times h + FP \times e + TN \times d + TN \times f \times g + TN \times f \times h + TN \times e)$
TPv	TP (validated)	$a \times TP + a \times FN + c \times g \times TP + c \times g \times FN$
TNv	TN (validated)	$d \times FP + d \times TN + e \times FP + e \times TN + f \times h \times FP + f \times h \times TN$
FPv	FP (validated)	$f \times g \times FP + f \times g \times TN$
FNv	FN (validated)	$b \times TP + b \times FN + c \times h \times TP + c \times h \times FN$
PP	PP (true positive + false positive)	TPv + FPv
PN	PN (true negative + false negative)	TNv + FNv
PPV	PPV = TP/PP = 1-FDR	TPv / PP
NPV	NPV = TN/PN = 1-FOR	TNv / PN
FNR	FNR = FN/P = 1-TPR	1 – TPR
FPR	FPR = FP/N = 1-TNR	1 – TNR
FOR	FOR = FN/PN = 1-NPV	1 – NPV
FDR	FDR = FP/PP = 1-PPV	1 – PPV
	Correlation (ϕ)	$(TPR \times TNR \times PPV \times NPV)^{.5} - (FNR \times FPR \times FOR \times FDR)^{.5}$

Appendix B

Summary of Clerical Review Procedure

For clerical review of pairs of linked records, compare the full name, address, date of birth, and gender for a survey record to the same fields for the prospective matched voter file record and tally the minor differences and major differences using the rules below. Match status is determined by calculating a match score: the number of major differences plus one-half the number of minor differences. If the match score is 2.5 or more, the match is considered incorrect, otherwise if the match score is less than 2.5, the match is considered correct.

This description is a summary. For the complete clerical review procedure, coder instructions, and results with ANES 2020 Time Series Study data, see Online Appendix 3.

Conditions for no difference:

- Identical records
- Capitalization, spacing, or apostrophes (e.g., DiMaggio = Di maggio; O'Brien = Obrien)
- Different city names should be ignored provided the ZIP codes are the same (e.g. "Stanford, CA 94305" and "Palo Alto, CA 94305" are considered the same)
- If a middle name is given in one record, using the initial in the other record is not considered a difference. E.g. "John Quincy Adams" and "John Q. Adams" are considered the same.
- Differences between substantively identical suffixes should be ignored. E.g., "Jr." = "Junior"
- In street names, differences in position or abbreviation of directional indicators (North, South, East, West) and abbreviation of street types (Avenue, Street, etc.) and numbers should be ignored. E.g., "East Fifth Avenue" and "E 5th Ave" and "5th Ave East" are considered the same.
- In days, months, and ZIP codes, if a leading zero is missing, this is no difference. E.g. 02/02/1982 = 2/2/1982, and for ZIP code, 04532 = 4532.

Minor differences:

- Any name or street field: a single typo
- First name: a recognizably appropriate nickname (e.g., Billy for William, or Peg for Margaret)
- Middle name: missing a middle name or initial when it was given in the other record
- Last name: omitting part of a hyphenated last name, when given in the other record

- Name suffix: omitting a name suffix such as Jr. or Sr., when given in the other record
- Unit number: omitting a unit number in the address, when given in the other record
- DOB: a partly missing date of birth (note “1” means missing for day and month)
- DOB: difference in one field (day, month, year) when the other two match

Major differences:

- First name different (more than a typo)
- Middle name different (not missing) if DOB also differs in month, day, or year
- Last name different (more than a typo)
- Last name extension different (e.g. Smith-Jones vs. Smith-White) (not missing)
- Name suffix (e.g. Junior or Senior) is different (not missing)
- Address differs in number, street name, unit number, state, ZIP, or a combination of these.
- DOB: all three fields are missing, or two or more fields are different, or a combination of missing and different with no matches; or one match, one different, and one missing.
- DOB differs by more than 15 years. This is cumulative with the previous difference, so “2/3/1950” compared to “8/26/1982” counts as 2 major differences.
- Name & DOB: a different first or last name combined with a DOB difference in at least two fields adds one major difference.
- Name & address: a different first or last name combined with a different street number, street name, or unit number adds one major difference.
- Name suffix & DOB: a different or missing name suffix (e.g., Jr in one record and missing in the other) combined with a DOB more than 15 years apart adds one major difference.