

COOPERATIVE CONGRESSIONAL ELECTION SURVEY

Vote Validation in the 2006 CCES

STEPHEN ANSOLABEHRE AND EITAN HERSH

Department of Government
Harvard University
1737 Cambridge Street
CGIS Knafel Building
Cambridge, MA 02138
617-496-0234
617-495-0438 (fax)
sda@gov.harvard.edu
hersh@fas.harvard.edu

October 7, 2008

Abstract

New technology and recent political reform have made vote validation an easier and more reliable process than it has been in the past. We present a basic summary of the vote validation procedure used in the 2006 CCES, a Web-based survey of nearly 35,000 Americans that has been validated electronically with new state-wide voter files. As the validation method in the CCES is quite different from the method used by the National Election Studies (NES) in the 1960s through 1980s, we compare the CCES procedure and results with the most recent midterm elections validated by the NES. We show that while the rate of vote misreporting is substantially higher in the 2006 Web-based survey, the pattern of misreporting is consistent with the NES samples. We also show how the large sample size in the CCES can be exploited to study phenomena beyond vote misreporting using the validated records.

1 Introduction

When analyzing public opinion surveys, scholars must usually take respondents at their word. When respondents claim that they earn \$50,000 a year, for instance, or that they have a bachelor's degree, or that they are very interested in politics, we have little choice but to believe them. When it comes to their reported voting behavior, however, we need not take respondents at their word because we can validate their answers with official election records. Validation is particularly useful for voting questions since the rate of reported voting in the United States far exceeds the true rate of voter turnout. Consider the 2004 Presidential election as an example. According to the Federal Election Commission, 56.7% of the voting age public cast a ballot in 2004.¹ In the National Election Study's (NES) 2004 post-election survey, however, 78.5% of respondents claimed to have voted, with a 95% confidence interval of $\pm 2.5\%$. Vote validation studies allow us to find out which respondents' vote records do not match their reported behavior and to theorize about why some people's claims about their vote history are at odds with the public record.

From 1964 to 1990, the National Election Studies validated the voting records of participants in ten surveys. These validation efforts have been well studied by political scientists and have led to three important findings: a.) misreporting is not random; it is systematically correlated with respondent characteristics, b.) misreporting is not only a function of the survey or validation process; it is also a function of respondents not telling the truth, and c.) misreporting can distort standard assumptions about who participates in American politics.

Since 1990, when the most recent NES validation was conducted, advances in technology and new government reforms have made validation an easier and more reliable process, yet no major national opinion survey has been validated since then. Our purpose here is to share some basic results of the validated 2006 Cooperative Congressional Election Survey (CCES), a Web-based study in which the vote reports of over 35,000 respondents were validated.

¹<http://www.fec.gov/pubrec/fe2004/federalections2004.pdf>

Because the NES has, until now, been the only game in town validating vote reports, we will compare the validation methods and findings of the NES and the CCES. Even though the validation and survey methods employed are quite different and even though sixteen years have elapsed since the most recent NES validation, the results from the two surveys reveal a remarkable degree of consistency. It turns out that misreporters in the 2006 study look much like misreporters in the NES studies.

2 The Lying Rate and the Validation Rate

Before exploring the data, let us consider the variables of interest in studies of vote validation. Most of the previous research on misreporting (e.g. Anderson and Silver 1986; Silver, Anderson, and Abramson, 1986; Belli et al. 2001; Fullerton, Dixon, and Borch 2007) has been concerned with explaining the misreporting, or lying, rate. This is the subset of non-voting respondents who report that they voted. The primary motivation of this line of inquiry is to figure out who is misreporting their vote and why. Other scholars (e.g. Bernstein, Chadha, and Montjoy 2001; Cassel 2003) are focused on a different variable, which we might refer to as the validation rate. This is the rate of actual voting among those respondents who claimed to have voted. The motivation for estimating the validation rate is to obtain a better measure of the true voting public. Briefly, let us explore how the lying rate and the validation rate stand in relation to the reported vote and the validated vote.

From a survey, we can observe an estimate of $Pr(R)$, the probability of a respondent reporting as having voted. Unless the survey is validated, we do not observe $Pr(V)$, the probability of respondents in the sample actually having voted. However, we do know that the pool of reported voters is made up of two groups: true voters who claim to have voted and non-voters who claim to have voted. To state this more formally,

$$Pr(R) = Pr(V)Pr(R|V) + Pr(R|\bar{V})(1 - Pr(\bar{V})) \quad (1)$$

Virtually no one who actually casts a ballot reports they did not vote. In the CCES

sample we discuss below, only $\frac{2}{5}$ of one percent of valid voters reported they did not vote, a rate of under-reporting consistent with NES validations (see Belli, Traugott, and Beckmann 2001, p. 483). Thus, for all intents and purposes we can assume that $Pr(R|V) = 1$.

The misreporting rate is equivalent to $Pr(R|\bar{V})$ in Equation 1 above. Let's refer to this variable as β . Given that $Pr(R|V) = 1$ and substituting β for $Pr(R|\bar{V})$, we can re-write equation 1 as $Pr(R) = \beta + (1 - \beta)Pr(V)$, and therefore $\beta = \frac{Pr(R) - Pr(V)}{1 - Pr(V)}$.

The validation rate, the probability of actually voting given one says he/she voted, is equivalent to $Pr(V|R)$, which we will call α . Using Bayes' Rule, we can see that the relationship between α and $Pr(R)$ is as follows:

$$Pr(R) = \frac{Pr(R|V)Pr(V)}{\alpha} \quad (2)$$

And, again, since $Pr(R|V)$ is effectively 1, $Pr(R) = \frac{Pr(V)}{\alpha}$. Therefore, quite simply, $\alpha = \frac{Pr(V)}{Pr(R)}$, the proportion of reported voters who actually voted. Estimating β helps us understand who among non-voters lie. Estimating α helps us understand who the real voters are.

Let us now observe how α and β contribute to the probability of reporting voting and the probability of actually voting.

$$Pr(R) = \frac{\beta}{1 - \alpha + \beta\alpha} \quad \text{and} \quad Pr(V) = \frac{\alpha\beta}{1 - \alpha(1 - \beta)} \quad (3)$$

These formal definitions of α and β , the lying rate and the validation rate, are useful in two ways. First, if α and β were consistent across samples and across elections, we could use these definitions to approximate the proportion of true voters in an unvalidated sample. For instance, in the CCES sample, β is about .7. Assume for a moment that this is a typical lying rate for a Web-based survey in a midterm election. If we draw a sample after an election in which the reported voting rate is .86, we could estimate using β the true turnout rate at .53. Second, these definitions lead to some useful predictions regarding the correlates of reported and validated voting. For example, suppose that $Pr(R)$ and $Pr(V)$ are both functions of

the same set of k explanatory variables plus some random noise. Knowing α , the probability that a vote reporter actually voted, we see that the true probability of voting is equal to $\alpha Pr(R)$. Since $Pr(R)$ and $Pr(V)$ are both functions of the same set of explanatory variables, we can predict that the coefficients on the x 's for $Pr(R)$ will be $\frac{1}{\alpha}$ times the coefficients on the x 's for $Pr(V)$. As it turns out, this relationship roughly fits the data. For the purpose of summarizing the data here, we will restrict our analysis to β . We will focus more on estimating α with the CCES and NES samples in a paper that is forthcoming.

3 Validation Methodology

The 2006 CCES is a Web-based survey, administered to respondents within two weeks of November election. Information on the sampling methodology can be found at CCES website.² All respondents were asked in a pre-election survey whether they were registered to vote. They were asked in a post-election survey if they had voted. Validators for the CCES sought records of all respondents.

The validation process used for the CCES is quite different from earlier validated studies conducted by the NES. The NES validations were conducted in person by interviewers who visited local election offices. Visiting the election offices in person was necessary because, during the period of the NES validations, not all records were kept electronically and voter lists were held only by local registrars. Since the last NES validation in 1990, both technology and the laws governing the maintenance of registration lists have changed. In 1993, Congress passed the National Voter Registration Act (NVRA), which included provisions detailing how voter records must be kept “accurate and current.” The Help America Vote Act (HAVA) of 2002 required every state (except North Dakota, which has no voter registration) to develop a “single, uniform, official, centralized, interactive computerized statewide voter registration list defined, maintained, and administered at the State level.”³ The 2006 election was the first election by which states were required to have their databases up and running.

²<http://web.mit.edu/polisci/portl/cces/commoncontent.html>

³http://www.fec.gov/hava/law_ext.txt

With records maintained at the local or county level rather than in a state-wide database and with digital records still quite new to registrars, it is probably a fair assumption that the quality of the record-keeping and record-updating was substantially worse in 1990 than it is today. As evidence of slow updating in decades past, consider the re-validation that the NES conducted of its 1988 validated survey. Two years after the initial validation, interviewers went back to see if records had changed. It turned out that 42.5% of all respondents thought to have been misreporting were now marked as having voted by the election office (Traugott, Traugott and Presser, 1992).

Even with centralization and digitization of voting records, there is still some variation across states in the quality of record-keeping. Because a validation process is only as good as the registration lists that are obtained from state agencies, in this study we only include respondents in states that maintain the most reliable lists. Polimetrix, Inc. conducted the validation process for the 2006 CCES. In its analysis of voting records, Polimetrix assigned each state a record quality rating of a 1, 2, or 3. A three indicates that records are kept very well in a particular state and the validators are quite confident that those whose records were not found or whose records were marked as having not voted did in fact misreport their participation. Our study focuses on the 26 states for which the records are of the highest quality.

The CCES survey was validated electronically. Polimetrix uses a “fuzzy matching” technique to search through state voter databases and match survey respondents to their voting records. “Fuzzy matching” involves using multiple pieces of information, in this case full name, age, and address to match respondents to voting records. For example, if a respondent identifies himself by a variation of his official name (e.g. Joe instead of Joseph), the matching process will use the other pieces of information to determine if it is a true match. If the probability of a true match is greater than 95%, Polimetrix consider the record matched. To the extent that the validation process contains error, most of the error seems likely to take the form of true voters whose names were not matched either because the voter lists were not updated by the state officials appropriately or because the information respondents

entered about their name and age was substantially different from the information contained in the voting records. There is a slighter chance of error in the other direction, of false voters being marked as true. This would be of the form of a husband and wife whose names and birthdays are nearly identical, one of whom was interviewed and did not vote, the other of whom was not interviewed and did vote. This scenario seems less likely but is still a possible instance of error.

4 Summary Statistics

In the 26 high quality states, 23,139 U.S. citizens in the CCES sample were asked whether they voted in the 2006 election. Of those, 18,283 answered the question yes or no. An additional 71 respondents answered that they did not know whether they had voted, and a significant number of respondents, over 5,500, did not answer the question at all. Of the 18,283 who gave an answer, 86.5% claimed to have voted, while 54.9% were validated as having voted. As for those who did not answer the voting question, one might expect that most of them did not vote. However, this group is divided almost evenly, with 45.9% having a validated vote. The voting behavior of these non-respondents stands in contrast to the small group of “I don’t know’s,” all but 9 of whom were validated non-voters.

What are we to make of this large number of people who did not answer the vote question but who did in fact vote? Some respondents probably accidentally skipped the vote question. They may have clicked the arrow leading to the next question on the survey without realizing they missed a prompt. Others, perhaps, were not comfortable revealing whether they had voted because voting is a private matter. Interestingly, about 1,000 of the roughly 5,500 respondents who did not answer the vote question *did* answer other questions related to their vote, such as how they voted (e.g. in person, by mail), for whom they voted in the Congressional race, and whether they had any registration problems when they went to vote. Among the 945 respondents who answered *all three* of these questions as if they had voted but did not answer the vote question, 68% are validated voters.

These survey questions related to voting can be used to improve the validation rate among reported voters as well. Whereas about 55% of respondents who did not leave the vote question blank are valid voters, 63.5% of people who answered the vote question and also gave answers to the three questions described above are validated as voters.

How does the validation rate in the CCES compare with the 1986 and 1990 NES validation studies? We compare the NES to these two years in the NES because they are the most recent midterm elections validated. As Cassel (2003) shows, over-reporting patterns in midterm elections are somewhat different than in Presidential years, at least in the NES samples. Thus, it is appropriate to compare the 2006 CCES survey only with other midterm years. In the 1986 NES, 52% of 2,072 respondents reported having voted, while 44% were validated as having voted. In the 1990 NES, 46% of 1,966 reported having voted, while 40% were validated as having voted. The percentage of misreporters, then, is four times larger in the CCES sample as in the NES sample.

Why is the rate of misreporting so much higher in the CCES sample than the NES samples? In a forthcoming paper, we take an in-depth look at sources of error on vote questions and we offer a more extended comparison between the CCES and NES samples. However, the short answer is that different survey formats (e.g. Internet, in-person, telephone) are prone to differing *levels* of misreporting but similar *patterns* of misreporting. That is, while the CCES Web-based survey elicited more misreporting than the NES, as we will see, the correlates of over-reporting are quite similar in the CCES and NES samples.

In Table 1, we provide basic summary statistics for each dataset as well as the conditional probabilities relevant to the discussion in Section 2 above. Note that respondents who did not answer the reported vote question are not included in this table.

In Table 1, we see first that while the valid vote rate in the CCES is 10-15 percentage points higher than the valid vote rates in the NES, the reported vote rate is 35-40 points higher. In all three samples, virtually all validated voters claim to have voted, as indicated by the third row of data. Perhaps the most striking difference between the CCES and NES samples is in the fifth row of data. This row represents the proportion of over-reporters

Table 1: Reported Vote and Validated Vote in the CCES and NES

| | NES (1986) | NES (1990) | CCES (2006) |
|----------------------------------|----------------------|----------------------|-----------------------|
| Reported Vote | 52.0% | 46.3 | 86.5 |
| Validated Vote | 44.0 | 39.9 | 54.9 |
| Pr(Report Vote Valid Vote) | 99.3 | 96.4 | 99.6 |
| Pr(Valid Vote Report Vote) | 84.2 | 83.0 | 63.2 |
| Pr(Report Vote Valid Not Vote) | 14.7 | 13.0 | 70.6 |
| Pr(Valid Not Vote Report Vote) | 15.8 | 16.9 | 36.8 |
| N (Vote Report) | 2072 | 1966 | 18283 |
| True Turnout Rate | 34.4 | 33.7 | 41.9 |

among non-voters, or β as we defined it above. In the CCES sample, 70% of non-voters claim to have voted, whereas in the NES sample only 31% or 15% claimed to have voted.

The final line of data in Table 1 requires further explanation. The numbers across the columns are not perfectly comparable. For the 2006 data, the true turnout rate of 41.9% is calculated by dividing the “maximum vote” in each state in the sample by the citizen voting age population (CVAP). Each state’s turnout rate is then weighted by its population to generate the statistic in the table. These data are taken from the U.S. Election Assistance Commission’s report on the 2006 election.⁴ Similar figures have not yet been calculated for the earlier elections. In the meantime, we have used the total votes cast for U.S. House races nationwide divided by the Voting Age Population for 1986 and 1990.

5 Who Overreports?

In spite of these stark differences in the rate of misreporting between the CCES and NES surveys, we find that similar respondents are misreporting in 2006 as in 1986 and 1990. In this section, we explore the correlates of misreporting.

The most common theory about the motivations of over-reporters is best summarized by Robert Bernstein et al. (2001, p. 24): “people who are under the most pressure to vote are the ones most likely to misrepresent their behavior when they fail to do so.” Two distinct

⁴http://www.eac.gov/clearinghouse/docs/eds-2006/edsr-final-adopted-version.pdf/attachment_download/file

categories of Americans seem to be under pressure to vote and therefore misreport when they abstain from voting: highly educated, politically engaged partisans and racial minorities or those living in areas dominated by racial minorities.

Across all validation studies, education is the most consistent predictor of over-reporting. Among those respondents who did not actually vote, it is the better educated ones who claim they did vote (Silver, et al., 1984; Bernstein et al. 2001; Belli, et al. 2001; Cassel 2003; Fullerton et al. 2007). In their recent paper comparing over-reporting about registration status versus over-reporting about vote status, Fullerton, Dixon, and Borch show that education is the only variable that is a strong predictor of over-reporting in both stages.

Aside from being more educated, over-reporters also tend to be more partisan, older, more likely to claim that they were contacted by a political party, more likely to be regular church attendees, and, as Belli, et al. (2001) explain, “similar to validated voters in that they see value in the political process.” This group of over-reporters consists of people who know and care about politics, and feel that it is a civic duty to vote.

Race has been the other widely cited correlate of over-reporting. Most scholars who have noticed the pattern of Blacks over-reporting (Belli et al. 2001; Fullerton et al. 2007; Duff, Hanmer, Park and White 2007) have theorized that in the wake of the Civil Rights movements, many Blacks feel duty-bound to vote, and are pressured to vote by their racial cohort. Given that the peak of the civil rights movement was in the 1950s and 1960s, it will be interesting to see whether the rate of over-reporting was higher in the 1970s and 1980s when most of the NES validation studies were conducted than in our 2006 survey.

While the relationship between race and over-reporting has been studied extensively, Bernstein et al. (2001, p. 28) explain that the relationship is dependent on racial context, and that the relationship is more complex than previously thought. In Congressional districts with higher concentrations of Blacks, Bernstein and his colleagues find that both Whites and Blacks over-report more than in White-dominated districts. In Congressional districts with high concentrations of Latinos, both Whites and Latinos over-report more than in White-dominated districts. Whites living in the Deep South tend to over-report their voting

behavior more than others as well. Fullerton and his research team add an additional layer of complexity in their recent finding that race is primarily a factor in over-reporting registration status and it plays less of a role in over-reporting one's vote record.

5.1 Model Specification

Using a probit model with the reported vote as dependent variable, we observe only those respondents in each sample who are not validated voters and examine who among them reported voting. We include as explanatory variables education, income, age, gender (male equals 1), church attendance, the number of years the respondent has lived in his/her current city of residence, marital status (1 if currently married, 0 otherwise), dummy variables for Hispanics, Blacks, and other non-Whites (Whites are the reference category), party identification (-1 for Democrats, 0 for Independents, 1 for Republicans), partisan strength (from pure independent to strong partisan), a dummy variable for respondents living in the South (former Confederate states) and an indicator of whether or not there was a Senate race (no senate race equals 1). These variables are among the most commonly used in models of vote over-reporting.

For each of the three samples, we also report the results of an alternative model that includes all of these variables in addition to two others: party contact and political interest. Party contact - whether or not the respondent claims that a political party contacted him/her in the run-up to the election - is often found in models of over-reporting (e.g. Bernstein et al. 2001; Cassel 2003; Fullerton et al. 2007). Although party contact is always found to be highly correlated with vote over-reporting, we suspect that there is endogeneity between the two variables. If we think that many misreporters lie about their voting record because they are politically engaged and feel they ought to have participated, then we should *not* expect all respondents who say they were in contact with a political party to be telling the truth about that either. We suspect that lying about voting is correlated with lying about other forms of political participation, including party contact. Nevertheless, because contact is often included in these models, we will show the results with it included as well.

The other addition to the second model is the political interest variable. We omit political interest in the first specification because in the CCES sample, a large number of respondents failed to answer the question. Of the approximately 22,900 respondents in the high-quality states, nearly 6,500 did not answer this question. Not wanting to lose so many observations, we report the results with and without this variable.

All non-dummy variables in Table 2 are standardized with a mean of zero and a standard deviation of one.

5.2 Results

In spite of differences in over-reporting rates between the NES surveys and the CCES survey, the variables are quite consistent. The first three rows of data show that in each of these samples, the older, better educated, and wealthier respondents among validated non-voters were most likely to report having voted. Looking at the lower portion of the table, we see that in all three samples, non-voters who identify as strong partisans, those who claim to be very interested in politics, and those who said that they were contacted by a party are all more likely to misreport than others. Notice that in each sample the models that include party contact and political interest are quite similar to those that do not. The coefficients on the other variables are attenuated in the (b) models, but the substantive story of the results does not change.

As for the remaining variables, the CCES is as inconsistent with each NES sample as the two NES samples are with each other. In all three samples, the coefficients on church attendance and gender are positive, but in the CCES the coefficients are larger and statistically significant. In the CCES and the 1986 NES, those who were longer-term residents in their towns were more likely to over-report than newer arrivals, but such a relationship is not evident in the 1990 sample. In the 1990 sample, non-voting Hispanics were more likely than Whites to report that they had voted, but this pattern did not emerge in 1986 or 2006. In the CCES and 1986 NES, Southerners were somewhat less likely to over-report, though only in the CCES is the coefficient statistically significant. In the 2006 CCES, respondents who

Table 2: Probit Model of Vote Over-Reporting

| Dep. Variable: Reported Vote | CCES '06 (1a) | CCES '06 (1b) | NES '90 (2a) | NES '90 (2b) | NES '86 (3a) | NES '86 (3b) |
|---------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| Education | 0.285*** (0.026) | 0.183*** (0.033) | 0.233*** (0.062) | 0.119* (0.067) | 0.268*** (0.068) | 0.141* (0.074) |
| Income | 0.178*** (0.025) | 0.089*** (0.032) | 0.185*** (0.070) | 0.178** (0.074) | 0.108 (0.076) | 0.115 (0.080) |
| Age | 0.370*** (0.025) | 0.302*** (0.033) | 0.219*** (0.061) | 0.156** (0.064) | 0.209*** (0.068) | 0.136* (0.072) |
| Male | 0.346*** (0.044) | 0.164*** (0.057) | 0.118 (0.106) | 0.025 (0.112) | 0.135 (0.112) | 0.019 (0.119) |
| Church Attnd. | 0.123*** (0.024) | 0.125*** (0.030) | 0.080 (0.057) | 0.093 (0.059) | 0.109* (0.057) | 0.087 (0.060) |
| Residence | 0.104*** (0.025) | 0.081** (0.032) | -0.007 (0.056) | 0.006 (0.059) | 0.190*** (0.067) | 0.198*** (0.071) |
| Married | -0.060 (0.047) | -0.054 (0.059) | -0.116 (0.113) | -0.139 (0.118) | -0.064 (0.121) | -0.019 (0.126) |
| Hispanic | -0.136* (0.078) | -0.003 (0.100) | 0.445*** (0.169) | 0.459*** (0.174) | 0.152 (0.215) | 0.177 (0.225) |
| Black | -0.082 (0.161) | -0.033 (0.196) | -0.129 (0.319) | -0.079 (0.330) | 0.418 (0.272) | 0.458 (0.286) |
| Other Non-White | 0.008 (0.095) | -0.110 (0.119) | -0.306 (0.315) | -0.290 (0.321) | 0.029 (0.355) | 0.026 (0.358) |
| Black*Income | -0.072 (0.054) | -0.042 (0.068) | 0.136 (0.094) | 0.127 (0.097) | 0.014 (0.087) | 0.009 (0.093) |
| Party ID | -0.010 (0.024) | -0.033 (0.029) | -0.081 (0.057) | -0.083 (0.060) | -0.083 (0.058) | -0.089 (0.061) |
| Partisan Strength | 0.204*** (0.022) | 0.155*** (0.028) | 0.102* (0.056) | 0.056 (0.058) | 0.157*** (0.058) | 0.082 (0.062) |
| South | -0.131*** (0.048) | -0.194*** (0.060) | 0.136 (0.127) | 0.107 (0.131) | -0.106 (0.120) | -0.136 (0.126) |
| No Senate | -0.195*** (0.048) | -0.196*** (0.060) | -0.076 (0.121) | -0.099 (0.126) | 0.014 (0.115) | -0.019 (0.120) |
| Political Interest | | 0.399*** (0.027) | | 0.357*** (0.063) | | 0.364*** (0.067) |
| Party Contact | | 0.446*** (0.055) | | 0.334** (0.151) | | 0.130 (0.139) |
| Constant | 0.979*** (0.051) | 0.793*** (0.075) | -1.103*** (0.127) | -1.115*** (0.135) | -1.019*** (0.126) | -1.020*** (0.133) |
| Observations | 5113 | 3518 | 1039 | 1028 | 917 | 893 |
| Log Likelihood | -2261 | -1438 | -376.6 | -350.2 | -344.8 | -319.4 |

*** p<0.01, ** p<0.05, * p<0.1

Standard errors in parentheses

lived in states holding Senate races were more likely to misreport than other respondents, but this was not the case in 1990 or 1986.

Notice that in addition to an indicator for Black respondents, we have also included an interaction between Black and income. This was done because the Black respondents in the CCES are not exactly representative of the Black population in general. The sampling algorithm Polimetrix uses ensures that a representative number of low income respondents and a representative number of Black respondents are included in the sample. However, the algorithm has a harder time gathering a representative sample of income by race. As such, the Black sub-sample is skewed toward the high end of the socio-economic spectrum. When the $\text{income} \times \text{Black}$ interaction term is included in the models, we see neither a main effect nor an interaction effect. If we left out the interaction term, however, we would see that in the NES samples, Black respondents over-reported more than Whites and in the CCES sample they over-reported less than Whites.

5.3 Contextual Variables

The large sample size of the CCES enables us to investigate contextual variables in ways not possible with the smaller NES samples. Using the NES, Bernstein et al. (2001) tested measures of racial context in their study of over-reporting, but they were forced to trade off important demographic variables in order to do so. With the CCES, such trade-offs are not necessary. We tested several models (not shown) which included as measures of racial context the percentage of the county of residence that is black and the percentage of county of residence that is Hispanic. These variables provide a more precise measure of racial context than was available to Bernstein, Chadha, and Montjoy, who used percentage of racial minorities in the Congressional district. For the most part, the relationships between over-reporting and racial context are insignificant. Hispanics living in counties with large Hispanic populations were slightly more likely to over-report than Hispanics elsewhere. Whites were slightly more prone to over-reporting in areas with more Whites than in areas with more Blacks.

We can also test the relationship between political context and misreporting using the CCES. In Table 3, we show a similar model to the one in Table 2, but we include measures of political context. We not only include an indicator of a Senate race, but also a measure of the competitiveness of the Senate race. Our measure of competitiveness is the vote share won by the winner of the race. Thus a value of .51 indicates a more competitive race than a value of .75. We include a parallel measure for the House race in the district, as well as an indicator of a Gubernatorial race.

Table 3 shows that the Senate race drives over-reporting among non-voters more than other kinds of political races. Respondents in states with Senate races were much more likely to misreport than other respondents, and among these respondents, those in states with very competitive races were even more likely to misreport. No such relationship exists with Gubernatorial races. For House districts, the coefficient is in the opposite direction from the Senate races. Non-voting respondents in less competitive districts were more likely to over-report than non-voting respondents in more competitive districts.

6 Validation Beyond Vote Report

Most scholarship on validation has focused on vote misreporting. However, validation can also reveal insights into other forms of participation. Fullerton et al. (2007) suggest incorporating registration into models of misreporting. They analyze separately the characteristics of vote over-reporters and registration over-reporters in the NES surveys. The CCES is superior to the NES for the purpose of studying validated registration because of the way registration and vote questions are asked in the two studies. In the NES, respondents are first asked if they voted in the recent election. Only if they answer ‘no’ does the interviewer ask them if they are registered. Additionally, in the more recent NES surveys, respondents who claimed not to be registered were not validated at all.

In the CCES, respondents were asked in the pre-election survey if they were registered to vote and were asked in the post-election survey whether they voted. The separation of

Table 3: Probit Model of Vote Over-Report with Political Context Variables

| Dep. Variable: Reported Vote | CCES '06 |
|---|----------------------|
| Education | 0.289*** (0.027) |
| Income | 0.193*** (0.026) |
| Age | 0.374*** (0.026) |
| Male | 0.338*** (0.045) |
| Church Attnd. | 0.114*** (0.024) |
| Residence | 0.106*** (0.026) |
| Married | -0.072 (0.048) |
| (Race Variables Included but Not Shown) | |
| Party ID | -0.005 (0.024) |
| Partisan Strength | 0.204*** (0.022) |
| South | -0.173*** (0.052) |
| No Senate Race | -2.453*** (0.368) |
| No Gov. Race | -0.055 (0.059) |
| Sen. Competitive | -1.038*** (0.168) |
| House Competitive | 0.043* (0.023) |
| Constant | 1.541*** (0.111) |
| Observations | 5036 |
| Log Likelihood | -2202 |

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

these questions allows for a more precise analysis of the correlates of over-reporting the two acts.

In Table 4, we show some basic information about registration validation in the CCES and the 1986 and 1990 NES samples. While the percentage of valid registrants is nearly identical in the three samples, almost all respondents in the CCES claimed to be registered, whereas fewer claimed so in the NES surveys. Valid registrants in the CCES sample were more likely to claim to have voted and to actually have voted than valid registrants in the NES samples as well.

Table 4: Reported Registration and Validated Registration in the CCES and NES

| | NES (1986) | NES (1990) | CCES (2006) |
|--|----------------------|----------------------|-----------------------|
| Reported Registration | 73.1% | 70.3 | 95.1 |
| Validated Registration | 65.3 | 61.5 | 62.5 |
| Pr(Report Vote Valid Registration) | 74.9 | 71.1 | 92.6 |
| Pr(Valid Vote Valid Registration) | 68.0 | 65.4 | 86.2 |
| Pr(Valid Registration Report Registration = 1) | 89.3 | 87.5 | 64.9 |
| Pr(Valid Registration Report Registration = 0) | ? | ? | 16.2 |
| N (Registration Report) | 2057 | 1948 | 22930 |

Notice the second to last row of data. We cannot calculate the cells in this row for the NES samples because the NES did not check the registration status of respondents who said they were not registered. But the CCES result in this row reveals an interesting difference between vote reporting and registration reporting. In the CCES sample, 98.3% of those claiming not to have voted were validated as non-voters. If respondents said they did not vote, by and large they did not vote. But registration status is different. Registration is surely a less memorable experience than voting. Many Americans probably do not remember or do not know their current registration status. As a result, 16% of reported non-registrants were in fact registered.

If some reported non-registrants in the NES samples were similarly registered, then the data in the second, third, and fourth rows of Table 4 would be incorrect. To see why, consider the row of validated registrants. In the 1986 NES, 553 out of 2,507 respondents reported not

being registered. If 16% of these respondents were actually registered, that would increase the percent of the total sample of valid registrants from 65.3% to 69.6%. Similarly, the NES 1990 value of 61.5% would increase to 66.4%. By the same logic, the two probabilities conditioned on valid registration would become smaller because the pool of valid registrants would be larger while the pool of reported and validated voters would stay the same. Thus, for example, the value of 71.1% in the third row of data would be reduced to 66.0%.

We can also exploit the large sample size in the CCES to reach beyond basic voting and registration validation and explore other political phenomena of interest. Consider Table 5. First, we look at respondents in states that have Election Day registration and compare them to respondents in states without Election Day registration. Among the 26 states in our sample, only two, Minnesota and Idaho, had Election Day registration for the 2006 election. When respondents in the pre-election survey were asked about their registration status, citizens of Minnesota and Idaho may not yet have been registered but then could have registered at the polls and voted. Thus, we should expect to see a higher rate of respondents reporting that they are not registered but being validated as registered in Minnesota and Idaho compared to the other states. This is exactly what we find. In Idaho and Minnesota, 31% of those who reported not being registered were validated as registered, a rate that is 16 percentage points higher than the rest of the sample. A difference of means test generates a p-value less than .01.

Table 5: Detailed Validation in the CCES

| | Election Day Reg. | No Election Day Reg. | |
|-----------------------------|----------------------------|-----------------------------|---------------------|
| $Pr(V_{reg} \bar{R}_{reg})$ | 31.0% | 15.3% | |
| $N(R_{reg} = 0)$ | 65 | 1055 | |
| | Election Day Voting | Vote by Mail | Early Voting |
| % of Valid Voters | 73.7% | 17.7% | 8.7% |
| $Pr(V_{vote})$ | 64.2% | 60.7% | 61.3% |
| N | 11955 | 3034 | 1471 |

In the second half of Table 5, we display the valid voting rates disaggregated by vote mode. For those respondents who claimed to have voted in the post-election survey, the CCES asked them if they voted on Election Day at the polls, if they voted by mail (including

by absentee ballot) or if they voted early. We might think that respondents are likely to misreport at different rates depending on their method of voting. Especially if memory plays a role in misreporting, as has been suggested (e.g. Belli et al. 1999), perhaps some voting experiences are more memorable than others. In our sample, 64.2% of reported voters who claimed to have voted the traditional way were telling the truth, a higher percentage than for voters who voted by mail or who voted early. The validation rates for early voting and vote by mail are statistically different from the traditional voting validation rate at the .05 level and .001 level, respectively. This suggests that misreporting may be more prevalent with the alternative vote modes than with traditional voting.

Surprisingly, about 19% of respondents who did not answer the vote question *did* answer the vote mode question. Recall that some 46% of those who did not answer the vote question ended up voting in the election. Among the 1,010 respondents who did not answer the vote question but did answer the vote mode question, 68% were validated. This rate of validation is even higher than among those who claimed that they were voters! That so many of the CCES respondents who did not answer the vote question actually voted suggests that scholars should be wary of suggestions that those who do not respond to vote questions ought to be treated as non-voters. This is probably especially true for Web-based surveys like the CCES.

Yet another interesting use of the validated survey is partisan voting patterns. Consider Tables 6 and 7. Here we list the states holding Senatorial and Gubernatorial elections in 2006. Next to each state, we show the percentage of reported voters who claimed to have supported the Democratic candidate, the percentage of validated voters who claimed to have supported the Democratic candidate, and the true vote share won by the Democratic candidate. Standard deviations are in parentheses. For the federal elections, the true vote share is taken from the Federal Election Commission. For the state elections, we used data from the *Washington Post*.

Table 6: Senatorial Democratic Vote Share by State

| Senatorial | Dem. Vote Share of Reported Voters (1) | N | Dem. Vote Share of Validated Voters (2) | N | Dem. Vote Share FEC (3) |
|--|--|-------|---|------|-------------------------------|
| California | 0.589 (0.492) | 2383 | 0.587 (0.492) | 1280 | 0.629 |
| Deleware | 0.618 (0.490) | 64 | 0.636 (0.486) | 43 | 0.710 |
| Florida | 0.611 (0.488) | 1381 | 0.615 (0.487) | 940 | 0.613 |
| Hawaii | 0.451 (0.502) | 47 | 0.467 (0.505) | 32 | 0.625 |
| Michigan | 0.636 (0.481) | 889 | 0.599 (0.490) | 752 | 0.580 |
| Minnesota | 0.527 (0.500) | 315 | 0.543 (0.499) | 373 | 0.605 |
| Missouri | 0.581 (0.493) | 594 | 0.557 (0.497) | 403 | 0.512 |
| Nebraska | 0.646 (0.480) | 113 | 0.648 (0.480) | 77 | 0.639 |
| Nevada | 0.446 (0.498) | 241 | 0.492 (0.501) | 166 | 0.425 |
| New York | 0.657 (0.475) | 1099 | 0.636 (0.482) | 706 | 0.684 |
| Ohio | 0.534 (0.499) | 787 | 0.566 (0.496) | 786 | 0.562 |
| Pennsylvania | 0.533 (0.499) | 1123 | 0.530 (0.499) | 802 | 0.587 |
| Tennessee | 0.501 (0.501) | 447 | 0.417 (0.494) | 167 | 0.486 |
| Washington | 0.583 (0.493) | 765 | 0.572 (0.495) | 507 | 0.587 |
| Total | 0.583 (0.493) | 10248 | 0.577 (0.494) | 7034 | 0.589 |
| $\frac{\sum_{j=1}^J n_j (Actual_j - Survey_j)^2}{N}$ | .002 (cols. 1 vs. 3) | | .002 (cols. 2 vs. 3) | | |
| $\frac{\sum_{j=1}^J n_j (Actual_j - Survey_j)}{N}$ | .017 (cols. 1 vs. 2) | | .033 (cols. 2 vs. 3) | | .035 (cols. 1 vs 3) |

Tables 6 and 7 enable us to analyze different components of measurement error. First, looking at columns 1 and 2, we can compare the Democratic vote share by reported voters with the Democratic vote share by validated voters. Significant differences between these two columns would suggest that partisanship is related to misreporting. An examination of the

two columns reveals that the estimates are quite similar. The Tennessee Senate race contains the biggest discrepancy between the reported voters and the validated voters, though even there a difference of means test fails to reject the null hypothesis that the two values are equal at the .05 level.

Comparing the second and third columns reveals error in the sampling frame. If the true Democratic vote share lies outside the 95% confidence interval of the validated Democratic vote share, then we conclude that the sample is not quite representative of the population of voters. Of the thirty-four elections listed here, seven of the true election results are outside the intervals of the estimates. These seven include the Senate races in Minnesota, New York, and Pennsylvania and the Gubernatorial races in Nebraska, New York, Pennsylvania, and Tennessee. Notice that in six of these seven cases, the CCES sample is biased in the Republican direction. Only in the Nebraska Gubernatorial race is the true vote significantly more Republican than the estimate from the validated sample.

Finally, the comparison between the reported vote rate and the true vote rate reveals the combination of error due to over-reporting and error due to sampling. At the bottom of Tables 6 and 7, we have calculated the mean squared error and absolute bias across columns. As is indicated, these calculation weight each state, j 's, sample by its number of observations, n . For the calculation of the bias (bottom row) between the reported vote and the validated vote, n_j refers to the number of observations for the reported column. Looking at the bias calculations, it is evident that there is slightly less bias between the validated column and the true column (column 2 versus column 3) as compared with the reported column and the true column (column 1 versus column 3), as we would expect.

7 Conclusion

Validation is essential if we are to appropriately interpret public opinion data in the presence of imperfect survey measurements, imperfect samples, and the “truthy” answers given by many survey respondents. We have shown that the same patterns of vote misreporting

identified nearly half a century ago persist to this day. New survey modes may lead to more or to less misreporting, but better educated, wealthier, older, more politically engaged non-voters will continue to misreport.

While we have identified several ways to use the vote validation results from the CCES, the validation process can extend even further beyond vote and registration behavior. One very useful direction for future research would be to validate other political behaviors, such as political financial contributions, and determine which survey respondents over-report those behaviors as well. Even more useful would be for other surveys to validate vote history and registration status. Given the consistency that was found in the correlates of misreporting between the 2006 CCES and the NES validation studies now two decades old, we suspect that there is a high degree of stability in misreporting. Nevertheless, without any contemporary studies being validated, we cannot know if the CCES is a typical sample with regard to the reported and validated vote.

Having here described the validation procedure and results from the CCES, we intend to use this data in service of three research projects we are now pursuing. First, we will continue to focus on how different sources of measurement error can be isolated and estimated with the help of validated surveys like the 2006 CCES. Second, we will develop a technique to sift out misreporters from true voters through the use of multiple questions about respondents' voting experiences. Finally, we will explore how validated surveys may alter our long-held assumptions about who in America votes.

References

- Anderson, B.A. and B.D. Silver. 1986. "Measurement and Mismeasurement of the Validity of the Self-Reported Vote." *American Journal of Political Science* 30(4):771–785.
- Belli, R.F., M.W. Traugott, M. Young and K.A. McGonagle. 1999. "Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring."
- Belli, R.F., M.W. Traugott and M.N. Beckmann. 2001. "What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics* 17(4):479–498.
- Bernstein, R., A. Chadha and R. Montjoy. 2001. "Overreporting Voting: Why It Happens and Why It Matters*." *Public Opinion Quarterly* 65(1):22–44.
- Cassel, C.A. 2003. "Overreporting And Electoral Participation Research." *American Politics Research* 31(1):81.
- Duff, B., M.J. Hanmer, W.H. Park and I.K. White. 2007. "Good Excuses: Understanding Who Votes With An Improved Turnout Question." *Public Opinion Quarterly* 71(1):67.
- Fullerton, A.S., J.C. Dixon and C. Borch. 2007. "Bringing Registration into Models of Vote Overreporting." *Public Opinion Quarterly* 71(4):649.
- Silver, B.D., B.A. Anderson and P.R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80(2):613–624.
- Traugott, M.W., SM Traugott and S. Presser. 1992. Revalidation of Self-Reported Vote. Technical report NES Technical Report Series: Doc. nes010160. Ann Arbor: American National Election Studies.

Table 7: Gubernatorial Democratic Vote Share by State

| Gubernatorial | Dem. Vote Share of Reported Voters | N | Dem. Vote Share of Validated Voters | N | Dem. Vote Share <i>Washington Post</i> |
|--|---------------------------------------|-------|--|------|---|
| Alaska | 0.441 (0.499) | 76 | 0.463 (0.503) | 47 | 0.456 |
| California | 0.415 (0.493) | 2277 | 0.417 (0.493) | 1237 | 0.411 |
| Colorado | 0.601 (0.490) | 440 | 0.608 (0.489) | 271 | 0.577 |
| Florida | 0.490 (0.500) | 1376 | 0.485 (0.500) | 929 | 0.464 |
| Georgia | 0.430 (0.495) | 665 | 0.415 (0.493) | 439 | 0.396 |
| Hawaii | 0.273 (0.449) | 48 | 0.350 (0.483) | 33 | 0.357 |
| Iowa | 0.573 (0.496) | 196 | 0.573 (0.496) | 138 | 0.551 |
| Idaho | 0.403 (0.492) | 143 | 0.406 (0.493) | 104 | 0.454 |
| Illinois | 0.538 (0.499) | 694 | 0.541 (0.499) | 604 | 0.556 |
| Michigan | 0.626 (0.484) | 902 | 0.591 (0.492) | 762 | 0.571 |
| Minnesota | 0.498 (0.501) | 309 | 0.499 (0.501) | 361 | 0.495 |
| Nebraska | 0.392 (0.490) | 115 | 0.403 (0.493) | 78 | 0.245 |
| Nevada | 0.477 (0.500) | 232 | 0.489 (0.501) | 160 | 0.478 |
| New York | 0.672 (0.470) | 1124 | 0.650 (0.477) | 721 | 0.704 |
| Ohio | 0.566 (0.496) | 771 | 0.597 (0.491) | 779 | 0.619 |
| Oklahoma | 0.648 (0.478) | 271 | 0.647 (0.479) | 148 | 0.663 |
| Pennsylvania | 0.543 (0.498) | 1131 | 0.525 (0.500) | 808 | 0.600 |
| South Carolina | 0.446 (0.498) | 275 | 0.420 (0.495) | 194 | 0.388 |
| South Dakota | 0.388 (0.490) | 89 | 0.399 (0.493) | 62 | 0.367 |
| Tennessee | 0.637 (0.481) | 443 | 0.615 (0.488) | 162 | 0.697 |
| Total | 0.523 (0.499) | 11577 | 0.522 (0.500) | 8037 | 0.502 |
| $\frac{\sum_{j=1}^J n_j (Actual_j - Survey_j)^2}{N}$ | | | | | |
| .002 (cols. 1 vs. 3) | | | .001 (cols. 2 vs. 3) | | |
| $\frac{\sum_{j=1}^J n_j (Actual_j - Survey_j)}{N}$ | | | | | |
| .013 (cols. 1 vs. 2) | | | .029 (cols. 2 vs. 3) | | .032 (cols. 1 vs 3) |