

DRAFT

American National Election Studies (ANES) 2004 Vote Validation: A Demonstration Exercise

www.electionstudies.org

Kelly Zidar
American National Election Studies (ANES)
Institute for Social Research
The University of Michigan

The ANES data utilized in this analysis were collected under a grant from the National Science Foundation (SES-0535334); the responsibility for any errors in analysis is the author's.

Special thanks are due to our collaborator, Dr. Michael McDonald of George Mason University, for his suggesting and seeking support for a new vote validation effort, for his leadership and guidance in the ongoing project, and for his work in obtaining the necessary data. Those interested in further validation efforts should contact Dr. McDonald directly.¹

¹ Dr. Michael P. McDonald
Associate Professor, George Mason University Non-Resident Senior Fellow, Brookings Institution
George Mason University
Department of Public and International Affairs
4400 University Drive – 3F4
Fairfax, VA 22030-4444
Office: 703-993-4191 Email: mmcdon@gmu.edu Web site: <http://elections.gmu.edu/>

Summary:

This document describes processes and results from a vote validation demonstration project using a California voter registration file and 2004 ANES records.²

Two randomly-assigned subgroups were created to assess matching procedures. Of 119 ANES California records with a pre and post-election interview, 75% (n=89) were matched by programmatically generating a list of candidate matches, with an optimal match manually selected from the candidate list. 25% of records (n=30) were matched using manual lookup procedures involving criteria similar to the programmatic match. For the manual match subgroup, a list of programmatically-selected candidate matches was produced to assess the comparability of match methods. The match success rate for respondents who claimed to be at least registered is 89.74% for the program match subgroup, and 92.86% for the manual match subgroup. Just over 95% of matched voters can be confirmed in the voting history as having voted, and all matched non-voters were confirmed as not having participated in the election. Detailed match and validation results are provided in sections 4 and 5.

1. Inputs

*California voter registration file:*³
16,497,785 records

*ANES 2004 file:*⁴
California records (n=138)
California records with both a pre and post-election interview (n=119)

2. Data preparation

California voter registration file

- File split into 20 text files
- Code written to:
 - Create 20 Access database files and import a single text file to each database;
 - Create 27 alphabetic Access database files (based on first character of last name, includes one database file for non A-Z characters);
 - Port data from the 1-20 database files to the appropriate alphabetic database file and add a file source (1-20) value;

² This exercise was performed to assess the viability of renewing self-reported vote validation for future ANES data collections. California records were chosen due to the perceived quality of California voter registration data and the availability of California records in the 2004 ANES data file. The findings reported herein should not be interpreted beyond a preliminary attempt to assess the feasibility of matching electronic records and a basic view of correspondence between ANES self-reported vote and voting history in the registration file for a small subset of participants in the 2004 ANES survey. Further information regarding the concept, history, and analysis of vote validation and the ANES project can be found in several publications available from the ANES Reference Library (http://www.electionstudies.org/resources/papers/reference_library.htm).

³ California Voter Registration File. 2004. [California Secretary of State, distributor and producer.] Sacramento, CA: California Secretary of State.

⁴ The American National Elections Studies (www.electionstudies.org). THE 2004 NATIONAL ELECTION STUDY [dataset]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor].

- Verify record counts after the 1-20 database files were reallocated to the alphabetic files;
- Sort the alphabetic database files according to last name, first name and additional primary name fields;
- Assign a unique id identifying the alphabetic database and record number; assign SAS-compatible field names.

ANES 2004 file

- Subset data to include California records with a pre and post-election interview (n=119).
- Performed conversions required to make ANES and registration values comparable.

3. Match process

Two randomly assigned subgroups were created to assess match procedures. 75% (n=89) of the 119 ANES records were assigned to the program match group, 25% (n=30) were assigned to the manual match group. Table 1 details the number of records in each subgroup for each category of variable V045018x, a summary of vote and registration status.

Table 1. Number of records by match subgroup and vote/registration status

	V045018x: Summary vote and registration status			
Randomly-assigned subgroup	R voter	R non-voter, registered	R non-voter, non-registered	<i>Total</i>
Program match	70	8	11	89 (75%)
Manual match	26	2	2	30 (25%)
<i>Total</i>	96	10	13	119 (100%)

Program match

89 program match records were processed through an automated query system where, for each record, the appropriate alphabetic database file was selected for further query based on the first letter of the respondent’s last name. The query returned records having an exact last name match where the first two characters of the respondent’s first name matched the same characters in either the first or middle name fields in the voter registration file. A virtual recordset was created to hold the records meeting query criteria. Each record in the virtual recordset was assigned a series of match indicators for key items; county, congressional district, zip, and gender indicators were assigned a dichotomous 0/1 value, with a 1 indicating a match between the ANES and voter registration record. Street address, name, date of birth, and phone number values were processed through a string match function⁵ to determine comparability on a 0-1 scale, with 1 assigned for an exact match. The indicators did not distinguish between non-matching valid values and values with missing data, therefore, two values with completely different character sets would receive the same zero result as 2 values with missing data or 1 value with missing data, all indicating a lack of a match for a particular value.

A composite match index was computed from the series of indicators, with additional weight given to an exact date of birth match and to a lesser extent an exact phone number match to

⁵ The string match function compared the type and number of characters in two strings but not the position of the characters within each string.

distinguish these records further. The top 25 records with the highest composite match index values were printed to a results file along with other summary information (the ANES record, the total number of records meeting the initial query criteria, colored highlighting for records with a matching date of birth and/or phone number) where a final match was selected from the candidate list of matches.

Records without a match were investigated in a manual second pass.

Manual match

A match for each of the 30 records was manually sought out in the appropriate alphabetic database file assessing comparability across the same key items utilized in the programmatic match.

After the manual match selection, all 30 records were processed through the automated query system in a programmatic second pass.

4. Match results

Overall match results

Table 2 describes match results by match method and voter type. Of the 89 records assigned to the program match subgroup, 78 respondents indicated that they were at least registered to vote and 70 of these records (89.74%) were matched in the voter registration file with one of the matches occurring in the manual second pass. The match success rate was much lower for the self-reported non-registered respondents in the program match subgroup, with a match found for only 1 of the 11 records (9.09%).

Of the 30 records assigned to the manual match subgroup, 28 respondents indicated that they were at least registered to vote and 26 of these records (92.86%) were matched in the voter registration file with one of the matches occurring in the programmatic second pass. A match was found for one of the two self-reported non-registered respondents in the manual match group.

When the 30 manual match subgroup records were processed through the automated query system after a final manual match was selected, all but 2 records were produced in candidate lists of matches, each having the highest composite match index within their candidate list. For the 2 records where a manual match was found and a program match was not, the first or last name in the survey data was missing a first or last character when compared to the values in the voter registration file.

Table 2. Match results by match method and voter type

Method/Voter Type (V045018x)	Total Records	Match First Pass	Match Second Pass	Match Success Rate
<i>Program</i>				
Voter/Registered ⁶	78	69	70	89.74%
Non-Registered	11	1	1	9.09%
<i>Manual</i>				
Voter/Registered	28	25	26	92.86%
Non-Registered	2	1	1	50.00%
<i>Total</i>	119	96	98	

Detailed match results

Table 3 illustrates detailed match results for all matched records and for match method subgroups.

Note: For ease of interpretation, the measures reflected in Table 3 have been transformed to a 0-100 scale. While exact matches for date of birth and phone were emphasized in the composite match index calculation in order to sort those records to the top of the candidate list of matches in the automated query system, this transformation is not included in the figures presented below. Instead, the percent of exact matches for date of birth and phone number are included in their respective string match columns in parentheses.

Overall, nearly 96% of the 98 records with a final match had equal values for county, 82% had corresponding values for congressional district, 87% had matching values for zip code, and just over 50% had equivalent values for gender (due to large amounts of missing data for gender in the registration file). The average string match function value for street address was 82.37, 80.82 for name (a high value considering middle name was available in the registration file but not the ANES file), 93.77 for date of birth, and 48.14 for phone. In total, nearly 87% of matched records had an exact match for date of birth, while approximately 34% had an exact match for phone number. The average composite match index was 77.68.

Generally match results are more favorable for the program match subgroup compared to the manual match subgroup. Any comparisons should be made with caution, however, due to sample size disparity between the two groups.

⁶ One of the respondents in this group without a matching voter registration record indicated in the survey that they were registered to vote in another state.

Table 3. Detailed match results for records with a final match

Method	n	(Mean) Comp Index	(%) County	(%) Cong District	(%) Zip	(%) Gender	(Mean) String Match: Street Address	(Mean) String Match: Name	(Mean) String Match: DOB^a	(Mean) String Match: Phone^a
<i>Overall</i>	98	77.68	95.92	81.63	86.73	52.04	82.37	80.82	93.77 (86.73%)	48.14 (33.67%)
<i>Program</i>	71	78.97	98.59	83.10	88.73	53.52	83.79	81.61	95.23 (88.73%)	47.18 (33.80%)
<i>Manual</i>	27	74.28	88.89	77.78	81.48	48.15	78.63	78.74	89.93 (81.48%)	50.67 (33.33%)

a. Percent of exact matches provided in parentheses

5. Validation results

Table 4 recounts the parity between respondent self-reported vote and registration status and the voting history from the voter registration file. For the 90 self-reported voters with a final match in the voter registration file (93.75% of all self-reported voters), 86 are represented in the voting history as having voted. This represents 95.56% of matched records and 89.58% of self-reported voters. All non-voting respondents regardless of registration status with a final match in the voter registration file were confirmed as not having participated in the election.

Table 4. Validation results by voter type

Voter Type (V045018x)	n	Matched n	Confirmed n (Match %, Total %)
Voter	96	90 (93.75%)	86 (95.56%, 89.58%)
Non-voter, registered	10	6 (60%)	6 (100%, 60%)
Non-voter, non registered	13	2 (15.39%)	2 (100%, 15.38%)
<i>Total</i>	119	98 (82.35%)	94 (95.92%, 78.99%)