Supplementary materials for “Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate”

Stephen Ansolabehere and Eitan Hersh

May 31, 2012

1 Variable Definitions and Summary Statistics

The CCES data used in this analysis come from the third release of the 2008 CCES. Analytic weights are utilized throughout the analysis, except in logit models (see Ansolabehere 2011). The validated vote variable is available in the public file. The other validated variables (registration, party, vote history, and race) are not available in the public file. The public variables and coding used in Section (3) are reported vote (CC403), education (V213, recoded 1=0, 2=1, 3 and 4=2, 5=3, 6=4), income (V246, recoded 1,2 and 3=1, 4 and 5 = 2, 6,7,8, and 9=3, 10,11,12,13,14=4, 15=missing), White(V211, if V211=1), Black (V211, if V211=2), other non-White (V211, if V211=something other than 1 and 2), married (V214, if V214=1), church attendance (V217, recoded 6=0, 4 and 5=1,3=2, 1 and 2=3, 7=missing), dummy variables for 5 age groups (using birthyear (V207)), ideological strength (V243, recoded as 3 and 6= 0, 2 and 4 = 1, 1 and 5 =), female (recoded from V208), political interest (V244, recoded 7 and 4 =0, 3 = 1, 2=2, 1=3), partisan strength (CC424, recoded 4,8=0, 3,5=1, 2,6=2, 1,7=3), recent mover (CC334, recoded 1,2,3 and 4=1, 5 and 6=0), and dummy variables for state of residence identified by the survey firm for post-election respondents. Note that these coding choices are employed to make the CCES variables and NES variables match as closely as possible.

The NES data used in this analysis come from the June 24, 2010 release of the cumulative data file. Data from 1980, 1984, and 1988 are used. The variables used are reported vote (VCF9151, recoded as 0 and 9=missing, 1=1, 5 and 8=0) and validated vote (VCF9155, recoded as 1=1, 3 and 5=0, other = missing. Note that if a validation was not possible (VCF9153 = 3 or 5), the validated vote variable is recoded as missing. Also note that in two of the election years under study self-reported non-registrants were not validated. Throughout this analysis these respondents are treated as validated non-voters. Thus the validated vote variable is replaced with a 0 if VCF9152 = 2. As independent variables, we use education (VCF0140a, recoded 1 and 2=0, 4 and 5=2, 6=3, 7=4, 8 and 9=missing), income (VCF0114 recoded 5=4, 0=missing), White (VCF0106a, recoded 0=missing, 1=1, other=0), Black (VCF0106a, recoded 0=missing, 2=1, other =0), other non-White (VCF0106a, recoded 0=missing 3,4,5, and 7=1, other=0), married (VCF0147, recoded 1=1, 2,3,4,5,7,8 =0, 9=missing), church attendance (VCF0130, recoded 5 and 7=0, 4=1, 3 and 2=2, 1=3, 8 and 9=missing), age groups recoded from VCF0101, ideological strength (VCF0803, recoded 9 and 4=0 2,3,5,6=1, 1 and 7=2, 0=missing), female (recoded from VCF0104), political interest (VCF0313, recoded 9 and 1=0, 2=1, 3=2, 4=3, 0=missing), partisan strength (VCF0301, recoded 0,4=0, 3,5=1, ,2,6=2, 1,7=3), and recent mover (VCF9001, recoded 0 and 1=1, 99=missing, other =0).

Table 7: Summary Statistics for variables in Section (2)

Variable	Obs.	Mean	St. Dev.	Min.	Max.
CCES					
Reported Turnout	26,256	0.84	0.37	0.00	1.00
Validated Turnout	31,735	0.64	0.48	0.00	1.00
Education	32,800	1.75	1.13	0.00	4.00
Income	30,531	2.91	1.01	1.00	4.00
White	32,800	0.74	0.44	0.00	1.00
Black	32,800	0.12	0.32	0.00	1.00
Other Non-White	32,800	0.14	0.35	0.00	1.00
Married	32,800	0.55	0.50	0.00	1.00
Church Attendance	32,449	1.52	1.12	0.00	3.00
Age 18-24	32,800	0.15	0.36	0.00	1.00
Age 25-34	32,800	0.17	0.38	0.00	1.00
Age 35-44	32,800	0.17	0.38	0.00	1.00
Age 45-54	32,800	0.21	0.41	0.00	1.00
Age 55 +	32,800	0.29	0.45	0.00	1.00
Ideological Strength	32,800	0.73	0.76	0.00	2.00
Female	32,800	0.52	0.50	0.00	1.00
Poly. Interest	32,800	2.21	0.97	0.00	3.00
Partisan Strength	26,161	2.02	1.10	0.00	3.00
Recent Mover	32,756	0.31	0.46	0.00	1.00
Match Confidence	29,004	0.83	0.03	0.64	0.89
Pct. Deadwood	25,705	0.04	0.04	0.00	0.38
Pct. Undeliverable	26,300	0.03	0.03	0.00	0.41
Vote History Discrep.	25,508	0.03	0.09	0.00	1.00
NES					
Reported Turnout	5,172	0.72	0.45	0.00	1.00
Validated Turnout	5,653	0.59	0.49	0.00	1.00
Education	5,856	1.51	1.14	0.00	4.00
Income	5,276	2.82	1.06	1.00	4.00
White	5,893	0.79	0.41	0.00	1.00
Black	5,893	0.12	0.33	0.00	1.00
Other Non-White	5,893	0.09	0.28	0.00	1.00
Married	5,889	0.57	0.49	0.00	1.00
Church Attendance	5,888	1.51	1.09	0.00	3.00
Age 18-24	5,886	0.12	0.33	0.00	1.00
Age 25-34	5,886	0.24	0.43	0.00	1.00
Age 35-44	5,886	0.20	0.40	0.00	1.00
Age 45-54	5,886	0.13	0.33	0.00	1.00
Age 55 +	5,886	0.31	0.46	0.00	1.00
Ideological Strength	5,829	0.50	0.57	0.00	2.00
Female	5,911	0.57	0.50	0.00	1.00
Poly. Interest	5,095	1.71	1.00	0.00	3.00
Partisan Strength	5,911	1.78	1.01	0.00	3.00
Recent Mover	5,888	0.10	0.30	0.00	1.00

2 Tables Replicated with Logit Models

Table 8: Logistic Regression Replication of Table 2

Dep Var: Reported Vote	CCES	NES
	2008	1980-1984-1988
Indep. Vars.:	$\hat{\beta}$	$\hat{\beta}$
Education	0.490**	0.478**
	(0.037)	(0.070)
Income	0.331**	0.342**
	(0.035)	(0.071)
Black	0.093	0.377*
	(0.145)	(0.175)
Other Non-Whte	-0.242**	-0.112
	(0.091)	(0.209)
Married	0.009	-0.311*
	(0.071)	(0.138)
Church Attendance	0.208**	0.315**
	(0.031)	(0.062)
Age 25-34	-0.382**	0.174
	(0.116)	(0.202)
Age 35-44	-0.123	0.517*
	(0.121)	(0.212)
Age 45-54	-0.466**	0.510*
	(0.119)	(0.245)
Age 55 +	0.268*	0.761**
	(0.121)	(0.214)
Ideological Strength	0.058	0.003
	(0.046)	(0.116)
Female	-0.907**	-0.053
	(0.068)	(0.132)
Poly. Interest	0.821**	0.427**
	(0.038)	(0.067)
Partisan Strength	0.423**	0.374**
	(0.030)	(0.068)
Recent Mover	-0.198**	-0.686**
	(0.071)	(0.217)
Year 1984		-0.077
		(0.159)
Year 1988		-0.163
		(0.161)
Constant	-3.414**	-4.327**
	(0.171)	(0.324)
Observations	6,380	1,633
Pseudo R^2	0.311	0.168
Log Likelihood	-2933	-793.0

Note: Standard errors are in parentheses. ** p<0.01, * p<0.05

Table 9: Logistic Regression Replication of Table 3

Dep Var: Reported Vote	Basic	Add State	Restricted to	Add Indiv.-Level
	Model	Fixed-Effects	Matched Rs	Conf. Measure
Indep. Vars.:	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
Education	0.490**	0.490**	0.548**	0.548**
	(0.037)	(0.038)	(0.050)	(0.051)
Income	0.331**	0.339**	0.316**	0.313**
	(0.035)	(0.036)	(0.047)	(0.047)
Black	0.093	0.142	0.290	0.286
	(0.145)	(0.149)	(0.198)	(0.198)
Other Non-Whte	-0.242**	-0.226*	-0.261	-0.258
	(0.091)	(0.097)	(0.135)	(0.135)
Married	0.009	0.007	-0.005	-0.003
	(0.071)	(0.072)	(0.093)	(0.094)
Church Attendance	0.208**	0.212**	0.236**	0.237**
	(0.031)	(0.032)	(0.041)	(0.041)
Age 25-34	-0.382**	-0.372**	-0.085	-0.090
	(0.116)	(0.118)	(0.166)	(0.166)
Age 35-44	-0.123	-0.150	-0.029	-0.034
	(0.121)	(0.123)	(0.170)	(0.170)
Age 45-54	-0.466**	-0.458**	-0.317	-0.329*
	(0.119)	(0.121)	(0.166)	(0.166)
Age 55 +	0.268*	0.243*	0.201	0.191
	(0.121)	(0.123)	(0.167)	(0.167)
Ideological Strength	0.058	0.063	0.001	0.001
	(0.046)	(0.047)	(0.061)	(0.061)
Female	-0.907**	-0.923**	-0.758**	-0.750**
	(0.068)	(0.069)	(0.089)	(0.089)
Poly. Interest	0.821**	0.817**	0.803**	0.802**
	(0.038)	(0.039)	(0.050)	(0.050)
Partisan Strength	0.423**	0.424**	0.433**	0.434**
	(0.030)	(0.031)	(0.040)	(0.040)
Recent Mover	-0.198**	-0.202**	0.085	0.078
	(0.071)	(0.073)	(0.094)	(0.094)
Confidence				-1.706
				(1.219)
Constant	-3.414**	-4.784**	-4.998**	-3.561*
	(0.171)	(0.939)	(1.484)	(1.801)
State Fixed-Effects?	No	Yes	Yes	Yes
Observations	6,380	6,380	3,710	3,710
Pseudo R^2	0.311	0.324	0.314	0.315
Log Likelihood	-2933	-2879	-1740	-1739

Note: Standard errors are in parentheses. ** p<0.01, * p<0.05

Table 10: Logistic Regression Replication of Table 4

Dep Var: Reported Vote	All	Best Quality	Next Best	2nd Worse	Worst
	Counties	Counties	Counties	Counties	Counties
Indep. Vars.:	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
Education	0.490**	0.444**	0.544**	0.487**	0.525**
	(0.037)	(0.073)	(0.078)	(0.081)	(0.073)
Income	0.330**	0.263**	0.262**	0.421**	0.424**
	(0.035)	(0.070)	(0.072)	(0.078)	(0.068)
Black	0.095	-0.350	-0.129	0.302	0.449
	(0.145)	(0.314)	(0.296)	(0.282)	(0.282)
Other Non-Whte	-0.245**	-0.058	-0.493**	-0.515*	0.005
	(0.092)	(0.164)	(0.184)	(0.216)	(0.194)
Married	0.006	-0.189	0.169	0.268	-0.187
	(0.071)	(0.143)	(0.146)	(0.151)	(0.137)
Church Attendance	0.206**	0.228**	0.127*	0.257**	0.220**
	(0.031)	(0.063)	(0.063)	(0.070)	(0.059)
Age 25-34	-0.374**	-0.350	-0.204	-0.575*	-0.426
	(0.116)	(0.226)	(0.226)	(0.252)	(0.234)
Age 35-44	-0.120	0.027	-0.200	-0.318	-0.073
	(0.121)	(0.241)	(0.245)	(0.255)	(0.241)
Age 45-54	-0.463**	-0.574*	-0.574*	-0.709**	-0.154
	(0.119)	(0.237)	(0.233)	(0.256)	(0.237)
Age 55 +	0.273*	0.153	0.122	0.196	0.503*
	(0.121)	(0.243)	(0.243)	(0.261)	(0.234)
Ideological Strength	0.059	0.051	-0.130	0.050	0.242**
	(0.046)	(0.095)	(0.094)	(0.099)	(0.088)
Female	-0.907**	-0.970**	-0.767**	-0.927**	-0.996**
	(0.068)	(0.135)	(0.136)	(0.147)	(0.130)
Poly. Interest	0.821**	0.856**	0.843**	0.841**	0.784**
	(0.038)	(0.080)	(0.077)	(0.081)	(0.071)
Partisan Strength	0.423**	0.420**	0.429**	0.398**	0.436**
	(0.030)	(0.062)	(0.061)	(0.064)	(0.059)
Recent Mover	-0.202**	-0.193	-0.234	-0.072	-0.317*
	(0.071)	(0.144)	(0.145)	(0.152)	(0.135)
Constant	-3.410**	-3.208**	-3.209**	-3.772**	-3.660**
	(0.171)	(0.344)	(0.334)	(0.373)	(0.335)
Observations	6,367	1,519	1,510	1,439	1,899
Pseudo R^2	0.311	0.294	0.297	0.340	0.335
Log Likelihood	-2928	-718.5	-713.4	-644.1	-823.2

Note: Standard errors are in parentheses. ** p<0.01, * p<0.05

Table 11: Logistic Regression Replication of Table 6

Dep Vars:	Vote	Registration	Absentee/Early	Party Affiliation
	Misreporting	Misreporting	Misreporting	Misreporting
Indep. Vars.:	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
Education	0.490**	0.526**	0.132**	-0.179**
	(0.037)	(0.049)	(0.026)	(0.069)
Income	0.331**	0.215**	-0.040	-0.056
	(0.035)	(0.043)	(0.032)	(0.081)
Black	0.093	-0.285	0.422**	0.550
	(0.145)	(0.177)	(0.104)	(0.339)
Other Non-Whte	-0.242**	-0.384**	-0.199	0.055
	(0.091)	(0.109)	(0.103)	(0.189)
Married	0.009	0.038	-0.065	0.025
	(0.071)	(0.088)	(0.060)	(0.151)
Church Attendance	0.208**	0.232**	0.034	0.149*
	(0.031)	(0.040)	(0.023)	(0.061)
Age 25-34	-0.382**	-0.176	-0.506**	0.756**
	(0.116)	(0.141)	(0.130)	(0.282)
Age 35-44	-0.123	-0.127	-0.348**	0.926**
	(0.121)	(0.148)	(0.126)	(0.288)
Age 45-54	-0.466**	-0.322*	-0.133	0.947**
	(0.119)	(0.145)	(0.121)	(0.287)
Age 55 +	0.268*	0.148	0.327**	0.765**
	(0.121)	(0.149)	(0.118)	(0.282)
Ideological Strength	0.058	0.025	-0.028	-0.208*
	(0.046)	(0.060)	(0.035)	(0.096)
Female	-0.907**	-1.007**	0.011	-0.209
	(0.068)	(0.086)	(0.053)	(0.139)
Poly. Interest	0.821**	0.673**	0.105*	-0.242**
	(0.038)	(0.043)	(0.042)	(0.088)
Partisan Strength	0.423**	0.365**	0.011	1.443**
	(0.030)	(0.037)	(0.027)	(0.075)
Recent Mover	-0.198**	-0.106	0.248**	0.142
	(0.071)	(0.089)	(0.067)	(0.159)
Constant	-3.414**	-1.976**	-2.218**	-3.030**
	(0.171)	(0.197)	(0.177)	(0.406)
Observations	6,380	4,552	12,515	1,797
Pseudo R^2	0.311	0.280	0.0207	0.278
Log Likelihood	-2933	-1905	-5182	-731.3

Note: Standard errors are in parentheses. ** p<0.01, * p<0.05