

Lies, Damn Lies, and Pre-Election Polling

Elias Walsh

University of Michigan, Ann Arbor,

`fgelias@umich.edu`,

Sarah Dolfin

Mathematica Research Inc.,

`SDolfin@mathematica-mpr.com`,

and John DiNardo,

University of Michigan, Ann Arbor and NBER

`jdinardo@umich.edu`

Corresponding Author: John DiNardo

Session Title: Polls, Votes, and Elections

Session Chair: JOHN DINARDO, University of Michigan

Discussants: MIREILLE JACOBSON, University of California-Irvine

HEATHER ROYER, University of California, Santa Barbara

Web Appendices:

1. Data appendix
2. Ten (10) web appendix tables.
3. Eleven (11) web appendix figures.
4. Five page discussion of intentions problem.
5. Poll questions

January 30, 2009

Web-Appendices

- 1) Appendix 1. Data Appendix.
- 2) Appendix 2. Ten (10) Web Tables.
- 3) Appendix 3. Eleven (11) Web Figures.
- 4) Appendix 4. Short Discussion of Intentions.
- 5) Appendix 5. Poll Questions.

Web Appendix: Data

Elias Walsh, Sarah Dolfin and John DiNardo

January 30, 2009

1 Election Poll Data

We include all general election polls including at minimum both of the major party candidates and completed after June 1 of the election year. We identify and drop polls reported multiple times. When a single poll reports responses to the question phrased to allow third party candidates and another question phrased to force a choice between the Democratic and Republican candidates we use only the poll that allows the respondent more options. When a poll reports the results of the full sample in addition to some number of subsamples we use only the sample that limits respondents to “likely voters.” Finally, we drop 39 polls with no reported sample size.

2 Election Results Data

We obtained official 1996, 2000 and 2004 presidential election results from the Federal Election Commission website: accessed <http://www.fec.gov/pubrec/fe1996/presge.htm> on February 11, 2008 accessed <http://www.fec.gov/pubrec/2000presgeresults.htm> on February 11, 2008 accessed <http://www.fec.gov/pubrec/fe2004/federalelections2004.pdf> on February 11, 2008 According to the FEC these results are “the official, certified federal election results obtained from each state’s election office and other official sources.” <http://www.fec.gov/pubrec/electionresults.shtml>. Official results of the 2008 presidential election were not yet available at the time of this writing. For this election we obtain results from the most up-to-date tallies from media websites or from the state Secretary of State office when available. These results are conveniently available with sources from Wikipedia.com (accessed from http://en.wikipedia.org/wiki/2008_presidential_election on November 19, 2008).

3 TESS Poll Data

The data and documentation for our survey, conducted by TESS and *Knowledge Networks* is available at <http://www.experimentcentral.org/data/data.php?pid=298>. The first wave was conducted between October 19th and October 24th, 2004. The second wave was conducted between October 26th and November 1st, 2004. We drop four observations from the Manski group with no response for probability of voting (three of these also have missing poll results). We also drop a combined 58 observations from both groups with missing poll results. The survey completion rate is 68% for the first wave and 71% for the second wave.

Web Appendix: Tables

Elias Walsh, Sarah Dolfin and John DiNardo

January 30, 2009

Web Appendix Table 1: November “Trial Heats” for 2000 U.S. Presidential Election

Of these 43 “last minute” national horse race polls from the 2000 U.S. Presidential Election only 3 of the 42 polls predicted either a “tie” or Gore ahead in the national race, despite the fact that the actual vote was a virtual “tie” (with Gore actually winning the popular vote). Consultation of the tables for the binomial distribution reveals that the probability of 42 or more “Bush” predictions out of the 45 displayed above is less than 5×10^{-7} percent. In making this calculation we use the assumption that Gore (the Democratic candidate) and Bush (the Republican candidate) received exactly the same number of votes, and the polls were independent samples.

Web Appendix Table 2: Descriptive Statistics of Pre-Election Poll Sample, 2000-2008

The implied sample size is calculated from the reported margin of error and a mean of 0.50. Similarly, the implied margin of error is calculated from the reported sample size and mean of 0.50. The differences between the reported and implied values can be attributed to rounding error in most (but not all) cases. The sample includes all available statewide pre-election polls completed on or after the first day of June in the election year. We drop 39 polls with missing sample size from all analyses. See text for a further discussion of the sample inclusion criteria. Over a third of all polls in our sample are conducted within two weeks of election day, and approximately 85% of polls are reported as polls of “likely voters” (as opposed to registered voters, adults, or no qualification at all). The intensity of polling by state tends to increase across the three election years, with a median (mean) of 9 (13.5) polls per state in 2008 and a median (mean) of 5 (10.1) polls per state in 2000.

Web Appendix Table 3: Total Percentage Reported in Polls

The poll totals in this table include all reported categories including undecided and other candidate respondents. The sum of the predicted shares in many polls do not add up to exactly 100 percentage points. Since nearly all polls report figures rounded to two digits, many of these sums can be explained by rounding error. We do observe a small fraction of polls that sum to an amount below that which can be explained by rounding error, although over 95% of the polls in our sample do add up to 99 percentage points or higher. In these cases, as in the case of rounding error, we handle the problem symmetrically to the undecided problem and use the share of the total reported poll as the prediction (see text for details).

Web Appendix Table 4: Descriptive Statistics for Undecideds and Other Candidates in Polls

“Conditional” shares are conditional on being having any undecided or ambiguous respondents (or third party, other or none in bottom panel). “Ambiguous” shares include categories that are lumped together, such as “Other/Undecided” as well as shares left unaccounted. The vote shares are the unweighted means across polls. Only about 1% of polls have no undecided or ambiguous respondents. In polls with undecided or ambiguous respondents these respondents account for approximately 7% of the total, most of whom are classified as undecided. The fraction of polls with third party candidates varies with the election year. In the 2000 election 3.7% of the electorate voted for a third party candidate, while only about 1% did so in 2004 or 2008. As might be expected, the 2000 polls included third party candidates (or other/none) over 90% of the time, while 2008 polls included these only about 70% of the time. The composition of the third party candidate components varies by election year.

Web Appendix Table 5: Pre-Election Polls

“Adj” means treating undecided respondents as strongly ignorable. The standardized prediction errors are calculated using the equation in the text. Under the null that the poll results are i.i.d. draws from the true distribution, the mean of the standardized prediction error is 0 and the variance is 1. Prediction errors and shares are in units of percentage points. See text for a discussion of this table.

Web Appendix Table 6: Pre-Election Polls, by Year

“Adj” means treating undecided respondents as strongly ignorable. The standardized prediction errors are calculated using the equation in the text. Under the null that the poll results are i.i.d. draws from the true distribution, the mean of the standardized prediction error is 0 and the variance is 1. Prediction errors and shares are in units of percentage points. The pattern of over-dispersion and bias is consistent across election years. The polls in 2004 are slightly less disperse and display the least bias of the three years. As noted in the text, the 2004 race was, to a large extent, a “replay” of the 2000 election, possibly making the 2004 election easier to predict. Indeed, use of the 2000 election result as a prediction would have correctly

guessed the winner 94% of the time: the polls we analyzed guessed the victor less than 74% of the time. The fact that most polls are conducted for “hard to predict races” only partly explains this fact, since even accounting for where polls are conducted, the 2000 election result will correctly guess the winner 83% of the time. In the 2000 and 2008 races the polls outperform this crude benchmark, but not by a large margin.

Web Appendix Table 7: Error in Pre-Election Polls, By Inclusion of Third Party Candidates

All columns treat undecided respondents as strongly ignorable. Under the null that the poll results are i.i.d. draws from the true distribution, the mean of the standardized prediction error is 0 and the variance is 1. Prediction errors and shares are in units of percentage points. Third party candidates received 1.3% of the popular vote in 2008, 1.0% in 2004 and 3.7% in 2000. In the 2000 elections polls that *included* any third party candidate provided forecasts with more bias for the Democratic candidate, less bias for the Republican candidate, and much less disperse forecasts for both. However, in 2004 we see precisely the opposite pattern.

Web Appendix Table 8: The Relation Between Forecast Errors and Prior Information

Each column is an OLS regression clustered by state. The dependent variable is the adjusted poll result, treating undecided respondents as strongly ignorable. See text for a discussion of the Democratic candidate results. The results are qualitatively similar for the Republican candidate results with somewhat less weight placed on the prior than for the Democratic candidate, though this difference is not statistically significant.

Web Appendix Table 9: Descriptive Statistics of Manski Poll

See text for a discussion of the TESS poll. The table demonstrates that the means of observed individual characteristics do not differ significantly within wave across the treatment and control groups with a p-value of 0.34 in wave 1 and 0.90 in wave 2. Approximately 85% of the control group sample responded that they intended to vote in the election. This fraction is statistically indistinguishable from the mean of the reported probability of voting in the Manski group sample. Over 75% of the Manski sample reported that they were virtually certain of going to the polls. A similar fraction also expressed certainty about their choice of candidate. With so few respondents expressing uncertainty about their voting behavior one might be surprised to see important differences in the estimated preferences of the experimental groups.

Web Appendix Table 10: Probabilistic vs. Usual Style Questions

Results pool both survey waves, employing DFL weights to account for differences in observed sample demographics between waves. In addition, we employ the survey weights provided by TESS designed to match the demographics of the surveyed sample to the U.S. Census and the Knowledge Networks Panel. The likely voter weights use the reported probability of voting (for the Manski group only) to adjust results. The missing data weights use DFL weights to account for 58 dropped observations with missing poll results on observed dimensions of demographics. Actual national 2004 election results were Bush 50.733%, Kerry 48.270%, and Other 0.996%. See text for discussion of pooled results. The results tabulated separately by wave do not demonstrate any significant differences between the Manski and the control group respondents.

Web Appendix Table 1: November “Trial Heats” for 2000 U.S. Presidential Election

Date	Size	Gore	Bush	Prediction	Polling Agency
11/3 - 5	1801	45.9	49.0	False	ABC News
11/2 - 4	1741	45.9	50.0	False	Poll
11/1 - 3	1495	46.9	50.0	False	
10/31 - 11/2	1280	45.9	49.0	False	
10/30 - 11/1	1032	46.4	50.5	False	
11/5 - 6	2350	46.0	48.0	False	Gallup/CNN
11/5 - 6	2350	† ^a 46.4	48.5	False	USA Today
11/4 - 5	2386	46.4	48.5	False	Poll
11/2 - 4	2733	44.8	50.0	False	
11/1 - 3	2222	45.3	49.5	False	
10/31 - 11/2	2128	44.2	50.5	False	
10/30 - 11/1	2123	45.3	49.5	False	
11/1 - 2	623	45.8	51.0	False	Marist College
11/3 - 5	1026	45.8	49.0	False	NBC News/Wall
11/2 - 3	751	45.4	48.5	False	Street Journal
10/31 - 11/2	808	46.2	48.4	False	Newsweek Poll
11/2 - 5	1301	47.0	49.0	False	Pew Research Center
11/2 - 5	1301	† 46.7	48.9	False	for the People & the
11/1 - 4	1307	46.2	49.5	False	Press Survey
11/4 - 6	1091	47.9	46.8	True	CBS News Poll
11/2 - 5	1273	44.7	48.9	False	
11/1 - 3	825	46.3	48.4	False	
11/1 - 4	1158	44.2	49.5	False	CBS News/New York Times Poll
11/1 - 2	1000	47.8	47.8	False	Fox News/Opinion Dynamics Poll
11/3 - 5	1348	47.0	47.0	True	The Harris Poll
11/1 - 5	§ ^b	44.4	46.5	False	ICR
11/5 - 6	1000	45.0	50.0	False	Tarrance Group-d-/
11/5 - 6	1000	† 45.6	51.1	False	Lake Snell Perry & Assoc.-R-
11/1-2,5	1000	41.6	51.7	False	Voter.com/
10/30 - 11/2	1000	41.6	51.7	False	Battleground Survey
10/29-31, 11/1	1000	43.3	51.1	False	
11/4 - 6	1292	46.0	47.9	False	Christian Science Monitor/
11/3 - 5	989	44.7	51.1	False	Investor’s Business Daily/
11/2 - 4	718	42.4	52.2	False	TIPP Poll
11/1 - 3	838	41.4	48.5	False	
10/31 - 11/2	1070	42.4	47.5	False	
10/30 - 11/1	1186	45.3	50.5	False	
11/3 - 5	1253	45.2	48.4	False	Hotline Bullseye Poll
10/31 - 11/2	1000	43.0	50.5	False	
11/4 - 6	1200	‡ ^c 48.0	46.0	True	Reuters/MSNBC
11/3 - 5	1200	‡ 46.0	47.0	False	Tracking Poll
11/2 - 4	1200	‡ 44.4	46.5	False	
11/1 - 3	1200	‡ 44.2	48.4	False	
10/30 - 11/2	1200	‡ 45.2	48.4	False	
10/29 - 11/1	1200	‡ 42.4	45.5	False	

^aThis poll is a duplicate of the one immediately above but applies allocation algorithm as if true allocated had not been reported. In principle, they should differ only because of rounding error.

^bNo sample size reported.

^cOnly “approximate” sample size reported

^eSource pollingreport.com.

Web Appendix Table 2: Descriptive Statistics of Pre-Election Poll Sample, 2000-2008

	All Polls	2000 Polls	2004 Polls	2008 Polls
Days before election	40.23 {39.01}	38.02 {41.71}	41.48 {40.08}	40.47 {35.77}
< two weeks before election	0.38	0.48	0.37	0.32
Poll of “likely voters”	0.84	0.85	0.83	0.83
Reported sample size	697.07 {280.32}	626.82 {213.26}	733.20 {276.62}	708.72 {314.96}
Reported margin of error	3.86 {0.61}	4.07 {0.57}	3.73 {0.56}	3.85 {0.64}
Implied sample size	703.76 {281.71}	620.31 {226.98}	743.96 {266.46}	715.54 {316.47}
Implied margin of error	3.88 {0.65}	4.04 {0.56}	3.74 {0.50}	3.90 {0.81}
Number of polls	1857	475	705	677
Number of races	143	47	46	50
Mean polls per race	12.99	10.11	15.33	13.54
Median polls per race	7	5	7.5	9
Minimum polls per race	1	1	1	1
Maximum polls per race	80	37	64	80

The implied sample size is calculated from the reported margin of error and a mean of 0.50. Similarly, the implied margin of error is calculated from the reported sample size and mean of 0.50. The differences between the reported and implied values can be attributed to rounding error in most (but not all) cases. The sample includes all available state-level pre-election polls completed on or after the first day of June in the election year. We drop 39 polls with missing sample size from all analyses. See text for a further discussion of the sample inclusion criteria. The source for all polls is pollingreport.com. Standard deviations in braces.

Web Appendix Table 3: Total Percentage Reported in Polls

	All Polls	2000 Polls	2004 Polls	2008 Polls
Mean	99.82	99.85	100.02	99.58
Standard Deviation	1.39	0.81	0.58	2.11
Minimum	81	89	92	81
5th percentile	99	99	99	98
10th percentile	99	99	100	99
25th percentile	100	100	100	100
90th percentile	101	100	101	101
95th percentile	101	101	101	101
Maximum	102	102	102	102
Number of polls	1857	475	705	677

Poll totals include all reported categories including undecided and other candidate respondents.

Web Appendix Table 4: Descriptive Statistics for Undecideds and Other Candidates in Polls

	All Polls	2000 Polls	2004 Polls	2008 Polls
Fraction of polls with any undecided or ambiguous	0.989	0.981	0.996	0.987
Share of poll (conditional)	0.074	0.092	0.064	0.073
Vote shares (conditional)				
Undecided	0.057	0.069	0.053	0.054
Ambiguous	0.014	0.021	0.010	0.013
Unaccounted	0.003	0.003	0.001	0.005
Fraction of polls with any third party, other or none	0.793	0.914	0.804	0.697
Share of poll (conditional)	0.033	0.053	0.023	0.028
Vote shares (conditional)				
Green	0.012	0.039	0.000	0.000
Independent	0.008	0.000	0.015	0.006
Libertarian	0.002	0.001	0.000	0.004
Reform	0.003	0.010	0.000	0.000
Constitution	0.000	0.000	0.000	0.000
Other	0.009	0.004	0.007	0.016
None	0.001	0.000	0.000	0.001

“Conditional” shares are conditional on having any undecided or ambiguous respondents (or third party, other or none in bottom panel). “Ambiguous” shares include categories that are lumped together, such as “Other/Undecided” as well as shares left unaccounted. Vote shares are the unweighted means across polls.

Web Appendix Table 5: Pre-Election Polls

	All Polls		“Likely Voters”		< 2 Weeks before Election	
	N = 1857		N = 1554		N = 704	
	Raw	Adj	Raw	Adj	Raw	Adj
Republican share	48.17		48.21		48.31	
	{6.12}		{5.90}		{5.36}	
Democratic share	49.99		49.98		49.75	
	{5.93}		{5.66}		{5.15}	
Predicted Republican	44.70	48.20	45.03	48.31	45.14	47.84
	{5.99}	{6.31}	{5.71}	{6.00}	{5.24}	{5.48}
Predicted Democratic	45.42	48.95	45.71	49.01	46.55	49.31
	{5.87}	{5.91}	{5.59}	{5.61}	{5.19}	{5.22}
Republican error	-3.48	0.03	-3.18	0.10	-3.17	-0.47
	{3.48}	{3.36}	{3.31}	{3.21}	{2.67}	{2.49}
Democratic error	-4.57	-1.04	-4.27	-0.96	-3.19	-0.43
	{4.00}	{3.45}	{3.79}	{3.29}	{3.02}	{2.70}
Standardized	-1.80	0.02	-1.63	0.07	-1.59	-0.22
Republican error	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
Variance of stand’d	3.32	3.07	2.82	2.69	1.86	1.58
Republican error	(0.16)	(0.16)	(0.12)	(0.14)	(0.12)	(0.10)
Standardized	-2.38	-0.55	-2.22	-0.51	-1.63	-0.23
Democratic error	(0.05)	(0.04)	(0.05)	(0.04)	(0.06)	(0.05)
Variance of stand’d	4.38	3.20	3.91	2.84	2.37	1.89
Democratic error	(0.19)	(0.14)	(0.20)	(0.13)	(0.15)	(0.12)
Republican victory	38.40		38.93		40.77	
Democratic victory	61.60		61.07		59.23	
Republican victory	40.01		40.22		38.64	
predicted						
Democratic victory	55.57		55.15		56.53	
predicted						
Mispredicted victor	20.73		20.46		19.18	
Mispredicted victor	24.23		24.26		28.41	
using prior race						
	One Observation Per Race					
	N = 143		N = 136		N = 117	
Republican share	50.01		49.68		50.11	
	{8.97}		{8.72}		{8.02}	
Democratic share	47.69		48.09		47.65	
	{8.92}		{8.53}		{7.85}	
Republican victory	53.15		52.21		53.85	
Democratic victory	46.85		47.79		46.15	
Mispredicted victor	16.08		16.18		19.66	
using prior race						

Web Appendix Table 6: Pre-Election Polls, by Year

	2000 Polls N = 475		2004 Polls N = 705		2008 Polls N = 677	
	Raw	Adj	Raw	Adj	Raw	Adj
Republican share	46.88		50.37		46.80	
	{6.43}		{4.90}		{6.39}	
Democratic share	49.37		48.69		51.78	
	{6.08}		{4.80}		{6.44}	
Predicted Republican	42.86	47.12	46.51	49.63	44.10	47.47
	{6.15}	{6.70}	{5.23}	{5.59}	{6.13}	{6.47}
Predicted Democratic	43.38	47.62	45.37	48.38	46.91	50.47
	{6.00}	{6.01}	{5.08}	{5.14}	{6.10}	{6.25}
Republican error	-4.02	0.24	-3.86	-0.74	-2.70	0.67
	{3.74}	{3.64}	{3.02}	{2.81}	{3.60}	{3.52}
Democratic error	-5.99	-1.75	-3.32	-0.31	-4.87	-1.31
	{4.54}	{4.07}	{3.02}	{2.71}	{4.11}	{3.53}
Standardized	-1.98	0.14	-2.07	-0.40	-1.39	0.37
Republican error	(0.09)	(0.08)	(0.06)	(0.06)	(0.07)	(0.08)
Variance of stand'd	3.47	3.34	2.86	2.45	3.43	3.23
Republican error	(0.27)	(0.33)	(0.27)	(0.26)	(0.23)	(0.19)
Standardized	-3.01	-0.90	-1.78	-0.17	-2.55	-0.70
Democratic error	(0.11)	(0.09)	(0.06)	(0.05)	(0.07)	(0.07)
Variance of stand'd	5.55	4.17	2.69	2.13	4.64	3.38
Democratic error	(0.49)	(0.31)	(0.17)	(0.16)	(0.28)	(0.21)
Republican victory	43.58		49.93		22.75	
Democratic victory	56.42		50.07		77.25	
Republican victory predicted	43.58		45.53		31.76	
Democratic victory predicted	52.84		47.52		65.88	
Mispredicted victor	19.58		26.95		15.07	
Mispredicted victor using prior race	26.95		12.91		34.12	
	One Observation Per Race					
	N = 47		N = 46		N = 50	
Republican share	49.90		52.36		47.97	
	{8.71}		{8.28}		{9.48}	
Democratic share	45.94		46.47		50.46	
	{8.32}		{8.28}		{9.50}	
Republican victory	57.45		58.70		44.00	
Democratic victory	42.55		41.30		56.00	
Mispredicted victor using prior race	23.40		6.52		18.00	

“Adj” means treating undecided respondents as strongly ignorable. The standardized prediction errors are calculated using the equation in the text. Under the null that the poll results are i.i.d. draws from the true distribution, the mean of the standardized prediction error is 0 and the variance is 1. Prediction errors and shares are in units of percentage points. Standard deviations in braces. Standard errors in parentheses. Standard errors on variance estimates are bootstrapped with 1000 repetitions.

Web Appendix Table 7: Error in Pre-Election Polls, by Inclusion of Third Party Candidates

	Republican Prediction Error			Democratic Prediction Error			Number of Polls
	Adj	Stand'd	Stand'd Var.	Adj	Stand'd	Stand'd Var.	
All 2000 polls	0.24 (0.17)	0.14 (0.19)	3.34 (0.33)	-1.75 (0.08)	-0.90 (0.09)	4.17 (0.32)	475
Buchanan included	-0.15 (0.20)	-0.07 (0.24)	2.77 (0.27)	-2.08 (0.10)	-1.06 (0.12)	4.03 (0.34)	292
Buchanan not included	0.87 (0.30)	0.47 (0.30)	4.08 (0.68)	-1.22 (0.15)	-0.64 (0.15)	4.32 (0.58)	183
Nader included	0.00 (0.16)	0.01 (0.20)	2.63 (0.26)	-2.03 (0.08)	-1.05 (0.10)	3.87 (0.32)	393
Nader not included	1.41 (0.56)	0.75 (0.50)	6.35 (1.27)	-0.40 (0.28)	-0.20 (0.25)	5.10 (0.94)	82
Both Buchanan and Nader included	-0.13 (0.20)	-0.06 (0.24)	2.74 (0.27)	-2.15 (0.10)	-1.10 (0.12)	3.96 (0.35)	277
Any third party candidate included	-0.09 (0.16)	-0.03 (0.19)	2.70 (0.25)	-1.88 (0.08)	-0.97 (0.10)	3.97 (0.32)	434
No third party candidate included	3.74 (0.80)	1.92 (0.74)	6.80 (1.66)	-0.38 (0.41)	-0.16 (0.38)	5.87 (1.62)	41
All 2004 polls	-0.74 (0.11)	-0.40 (0.10)	2.45 (0.26)	-0.31 (0.06)	-0.17 (0.05)	2.13 (0.15)	705
Nader included	-0.76 (0.14)	-0.42 (0.13)	2.54 (0.34)	-0.79 (0.08)	-0.44 (0.07)	2.08 (0.18)	391
Nader not included	-0.72 (0.16)	-0.38 (0.15)	2.34 (0.37)	0.29 (0.09)	0.16 (0.08)	2.00 (0.25)	314
Any third party candidate included	-0.92 (0.12)	-0.51 (0.11)	2.57 (0.31)	-0.57 (0.07)	-0.30 (0.06)	2.16 (0.19)	567
No third party candidate included	-0.03 (0.21)	0.02 (0.21)	1.72 (0.22)	0.75 (0.11)	0.38 (0.11)	1.67 (0.21)	138
All 2008 polls	0.67 (0.14)	0.37 (0.14)	3.23 (0.19)	-1.31 (0.07)	-0.70 (0.07)	3.38 (0.20)	677
Any third party candidate included	0.04 (0.15)	0.07 (0.16)	2.93 (0.23)	-1.58 (0.08)	-0.87 (0.08)	3.33 (0.27)	472
No third party candidate included	2.13 (0.25)	1.05 (0.26)	3.26 (0.32)	-0.68 (0.13)	-0.31 (0.13)	3.30 (0.32)	205

All columns treat undecided respondents as strongly ignorable. Under the null that the poll results are i.i.d. draws from the true distribution, the mean of the standardized prediction error is 0 and the variance is 1. Prediction errors and shares are in units of percentage points. Third party candidates received 1.3% of the popular vote in 2008, 1.0% in 2004 and 3.7% in 2000. Standard errors in parentheses. Standard errors on variances are bootstrapped with 1000 repetitions.

Web Appendix Table 8: The Relation Between Forecast Errors and Prior Information

	Dependent Variable = 2008 Polls							
	Republican Candidate				Democratic Candidate			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
2008 Outcome	0.861 (0.041)		0.569 (0.091)	0.571 (0.104)	0.821 (0.041)		0.507 (0.085)	0.492 (0.099)
2004 Outcome		0.899 (0.053)	0.338 (0.105)	0.440 (0.215)		0.855 (0.045)	0.360 (0.090)	0.500 (0.154)
2000 Outcome				-0.205 (0.114)				-0.144 (0.106)
1996 Outcome				0.130 (0.076)				0.023 (0.135)
Constant	7.166 (1.978)	0.908 (2.595)	3.289 (2.159)	2.716 (2.260)	7.967 (2.098)	10.108 (2.250)	7.222 (1.756)	7.007 (2.591)
R-squared N = 677	0.723	0.690	0.738	0.741	0.715	0.692	0.733	0.736

	Dependent Variable = 2004 Polls							
	Republican Candidate				Democratic Candidate			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
2004 Outcome	0.986 (0.032)		0.927 (0.122)	0.951 (0.131)	0.915 (0.032)		0.886 (0.099)	0.881 (0.104)
2000 Outcome		0.927 (0.089)	0.061 (0.119)	0.143 (0.093)		0.828 (0.111)	0.033 (0.103)	0.006 (0.128)
1996 Outcome				-0.139 (0.159)				0.043 (0.137)
Constant	-0.034 (1.567)	5.125 (4.268)	-0.006 (1.616)	0.456 (1.540)	3.851 (1.643)	8.480 (5.447)	3.666 (1.700)	3.095 (2.472)
R-squared N = 705	0.747	0.681	0.747	0.750	0.729	0.582	0.730	0.730

	Dependent Variable = 2000 Polls					
	Republican Candidate			Democratic Candidate		
	(1)	(2)	(3)	(1)	(2)	(3)
2000 Outcome	0.883 (0.049)		0.745 (0.081)	0.764 (0.047)		0.594 (0.143)
1996 Outcome		1.021 (0.072)	0.185 (0.093)		0.932 (0.059)	0.228 (0.159)
Constant	5.719 (2.213)	6.574 (2.726)	4.834 (2.298)	9.920 (2.399)	1.090 (3.067)	6.889 (2.467)
R-squared N = 475	0.717	0.637	0.721	0.598	0.558	0.602

Web Appendix Table 9: Descriptive Statistics of Manski Poll

	Wave 1		Wave 2	
	Probabilistic	Control	Probabilistic	Control
Number of respondents	647	682	675	711
Fraction expressing no uncertainty in candidate preference	0.764		0.767	
Fraction expressing little (<10%) uncertainty in candidate preference	0.897		0.908	
Probability of voting				
Mean	0.841	0.839	0.857	0.857
Standard deviation	0.338	0.368	0.315	0.351
10th percentile	0	0	0.2	0
25th percentile	0.99	1	0.99	1
50th percentile	1	1	1	1
Demographics				
Age	47.209	47.443	47.108	47.498
{Standard deviation}	{16.908}	{16.744}	{16.940}	{17.701}
White	0.810	0.792	0.796	0.788
Male	0.488	0.493	0.484	0.498
Household head	0.819	0.833	0.839	0.826
Married	0.603	0.589	0.582	0.536
Metro area	0.807	0.826	0.847	0.840
Employed	0.621	0.572	0.573	0.589
Less than high school	0.130	0.166	0.166	0.166
High school graduate	0.272	0.224	0.273	0.276
Some college or associate degree	0.332	0.359	0.289	0.293
B.A. or higher	0.266	0.251	0.273	0.266
Northeast	0.176	0.188	0.188	0.173
Midwest	0.283	0.249	0.276	0.294
South	0.331	0.331	0.313	0.329
West	0.210	0.232	0.224	0.204
F-statistic from joint test of significance	1.12		0.54	
p-value from joint test of significance	0.3393		0.8987	

Standard deviations in braces.

Web Appendix Table 10: Probabilistic vs. Usual Style Questions

	Probabilistic Group			Control Group		
	Bush	Kerry	Other	Bush	Kerry	Other
Wave 1	N = 647			N = 682		
Survey weighted	46.534 (2.137)	49.873 (2.157)	3.593 (0.751)	47.190 (2.276)	49.551 (2.290)	3.259 (0.823)
P(vote) > 0	N = 577			N = 572		
Survey weighted	48.919 (2.318)	48.078 (2.336)	3.004 (0.750)	48.806 (2.487)	48.900 (2.496)	2.293 (0.840)
Above, and participation weighted	48.915 (2.403)	48.410 (2.415)	2.675 (0.637)			
Above, and missing data weighted	48.585 (2.437)	48.761 (2.455)	2.654 (0.628)	48.763 (2.490)	48.949 (2.499)	2.288 (0.839)
p-values						
Bush(M=1) = Bush(M=0)		0.9593				
Kerry(M=1) = Kerry(M=0)		0.9573				
Joint		0.9445				
Wave 2	N = 675			N = 711		
Survey weighted	45.528 (2.069)	50.997 (2.061)	3.474 (0.661)	45.519 (2.144)	49.110 (2.153)	5.371 (1.093)
P(vote) > 0	N = 613			N = 609		
Survey weighted	45.037 (2.173)	52.117 (2.173)	2.846 (0.647)	47.435 (2.337)	49.507 (2.341)	3.058 (0.931)
Above, and participation weighted	44.913 (2.232)	52.425 (2.231)	2.662 (0.661)			
Above, and missing data weighted	44.772 (2.237)	52.567 (2.238)	2.661 (0.656)	47.408 (2.338)	49.551 (2.342)	3.042 (0.924)
p-values						
Bush(M=1) = Bush(M=0)		0.4155				
Kerry(M=1) = Kerry(M=0)		0.3518				
Joint		0.6374				
Wave 1 & 2 Combined	N = 1322			N = 1393		
Survey weighted	46.037 (1.485)	50.429 (1.490)	3.534 (0.501)	46.364 (1.563)	49.333 (1.573)	4.303 (0.684)
P(vote) > 0	N = 1190			N = 1181		
Survey weighted	46.973 (1.582)	50.102 (1.589)	2.925 (0.495)	48.119 (1.705)	49.204 (1.709)	2.676 (0.627)
Above, and participation weighted	46.886 (1.633)	50.445 (1.636)	2.669 (0.459)			
Above, and missing data weighted	46.655 (1.646)	50.687 (1.652)	2.657 (0.454)	48.084 (1.706)	49.250 (1.711)	2.666 (0.624)
p-values						
Bush(T=1) = Bush(T=0)		0.5467				
Kerry(T=1) = Kerry(T=0)		0.5457				
Joint		0.8295				

In all results we employ survey weights provided by TESS designed to match the demographics of the surveyed sample to the U.S. Census and the Knowledge Networks Panel. The pooled results employ DFL weights to account for differences in observed sample demographics between waves. Likely voter weights use the reported probability of voting (for the Probabilistic group only) to adjust results. The missing data weights use DFL weights to account for 58 dropped observations with missing poll results on observed dimensions of demographics. All weights (except the TESS survey weights) are estimated using probit regressions of the appropriate outcome on a flexible set of the individual demographics including age, age squared, and dummies for each of the categorical variables in web appendix Table 9. Actual national 2004 election results were Bush 50.733%, Kerry 48.270%, and Other 0.996%. Heteroskedasticity robust standard errors in parentheses.

Web Appendix: Figures

Elias Walsh, Sarah Dolfin and John DiNardo

January 30, 2009

Web Appendix Figure 1: Density Estimates of Standardized Prediction Errors, by Election Year

The figure displays a kernel density of the standardized prediction errors for presidential state races by election year. The vertical lines are the estimated mean associated with the appropriate density. In comparison to the standard normal density, the theoretical prediction under random probability sampling, the poll densities are more disperse and are not centered at 0, indicating bias. The bandwidth for density estimation was chosen by ocular inspection and is 0.2.

Web Appendix Figure 2: Density Estimates of Standardized Prediction Errors, by Poll Subgroup

The figure displays a kernel density of the standardized prediction errors for presidential state races by poll subgroup. In comparison to the standard normal density, the theoretical prediction under random probability sampling, the poll densities are more disperse, though the polls within two weeks of the election do show less dispersion. The bandwidth for density estimation was chosen by ocular inspection and is 0.2.

Web Appendix Figures 3 & 4: The Relation Between Forecasts and Election Results

Each circle represents the mean of all poll results in a statewide election. The dashed line is the estimated line from a regression of the poll prediction on the actual election outcome. The solid line is the 45-degree line. The slope of the estimated line is always less than 1 (see also web appendix Table 8). Thus bias in polls tends to work in a way that understates larger vote shares and overstates smaller vote shares. This could be explained as a result of “honest Bayesian” type behavior on the part pollsters, or simply an artifact of other problems in polling that cause bias. For Democratic candidates the point at which the regression line crosses the 45-degree line is below 50%, while for Republicans this crossing point tends to be higher. If pollsters do act like “honest Bayesians” then these crossing points may be indicative of the pollsters’ prior beliefs about a candidate’s vote share. If for example, pollsters are reporting the maximum posterior density, then the nonzero intercept and departure of the slope from 1 are the consequence of the standard omitted variable calculations where the omitted variable is pollsters’ prior information. These general findings are not changed much if we limit the analysis to only those polls conducted within two weeks of the election.

Web Appendix Figures 5 & 6: The Relation Between Forecast Errors and Election Results

Figures 5 and 6 are analogues to Figures 3 and 4 with poll prediction errors in the place of the predicted vote shares. Each circle represents the mean of all poll prediction errors in a statewide election. The estimated line from a regression of the poll prediction errors on the actual election outcome is always negative. The main benefit to displaying the prediction errors rather than the predicted shares is that the scatter plot is more clearly presented.

Web Appendix Figure 7: Distribution of Polls Across States By Election Result and Number of Electoral Votes

Each circle represents a statewide election. The area of the circle is proportional to the number of polls in that race. Races with more polls tend to be concentrated in states with more electoral votes and in states that are more highly contested. We would expect to see very large circles in states that both have many electoral votes and are close races, however, the only state with more than 40 electoral votes is California, a state that is not particularly competitive.

Web Appendix Figure 8: Standardized Prediction Errors Over Time

The figure displays scatter plots of standardized prediction errors for presidential statewide races and quantile

regressions at the 10th, 50th, and 90th quantile. The lines in panels (a) and (c) present the results of a quantile regression of the prediction errors on the number of days before the election and a constant term. Panels (b) and (d) present the 10th and 90th quantiles, and associated confidence intervals from a design-adaptive bandwidth quantile regression, limiting the sample to only those polls within 10 weeks of the election. The two dotted horizontal lines in each of panels (b) and (d) indicate the theoretical prediction of the 90th and 10th percentiles under standard normality (1.28/-1.28). The panels demonstrate that dispersion in the poll errors diminishes over time, but even for the closest polls to the election the dispersion exceeds that of a standard normal density.

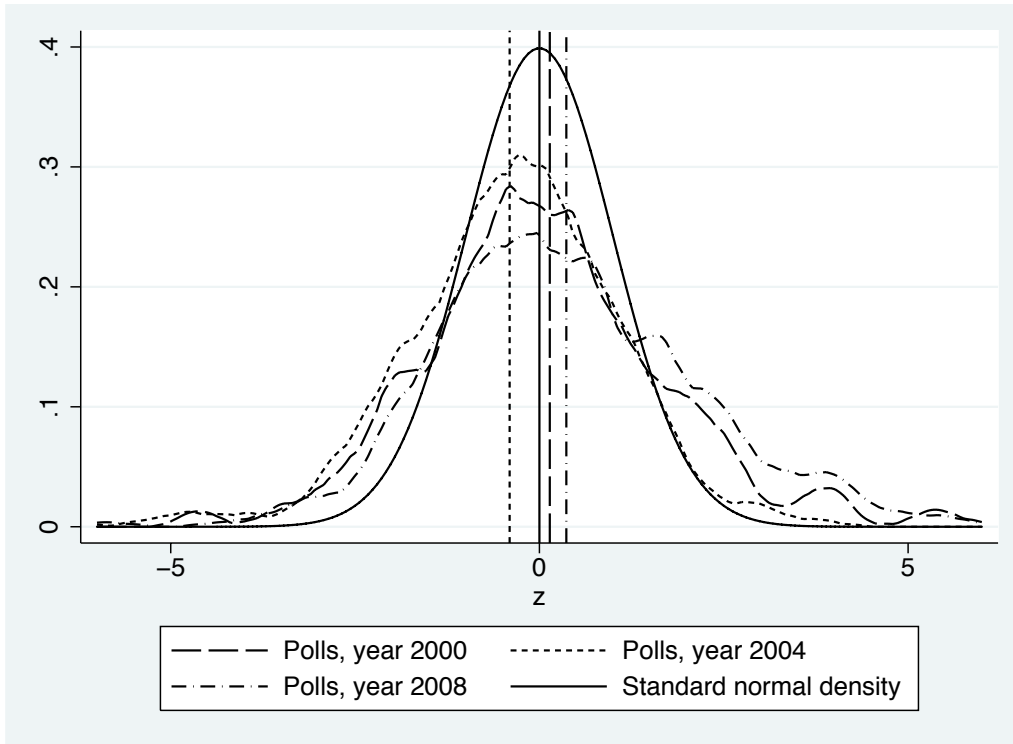
Web Appendix Figures 9 & 10: Standardized Prediction Errors Over Time, by Election Year

The figures display scatter plots of standardized prediction errors for presidential statewide races and quantile regressions at the 10th, 50th, and 90th quantile for polls separately by election year. The lines in panels (a), (c), and (e) present the results of a quantile regression of the prediction errors on the number of days before the election and a constant term. Panels (b), (d), and (f) present the 10th, and 90th quantiles and associated confidence intervals from a design-adaptive bandwidth quantile regression, limiting the sample to only those polls within 10 weeks of the election. The two dotted horizontal lines in each of panels (b), (d), and (f) indicate the theoretical prediction of the 90th and 10th percentiles under standard normality (1.28/-1.28). As in Figure 9, the panels generally demonstrate that dispersion in the poll errors diminishes over time, but even for the closest polls to the election the dispersion exceeds that of a standard normal density. We see some variation across election, with the 2004 polls for both the Republican and Democratic candidate displaying more-or-less constant dispersion over time. Also, the design-adaptive bandwidth quantile regressions do not always reject the prediction for the 10th and 90th quantiles of the standard normal density for the closest polls to the election.

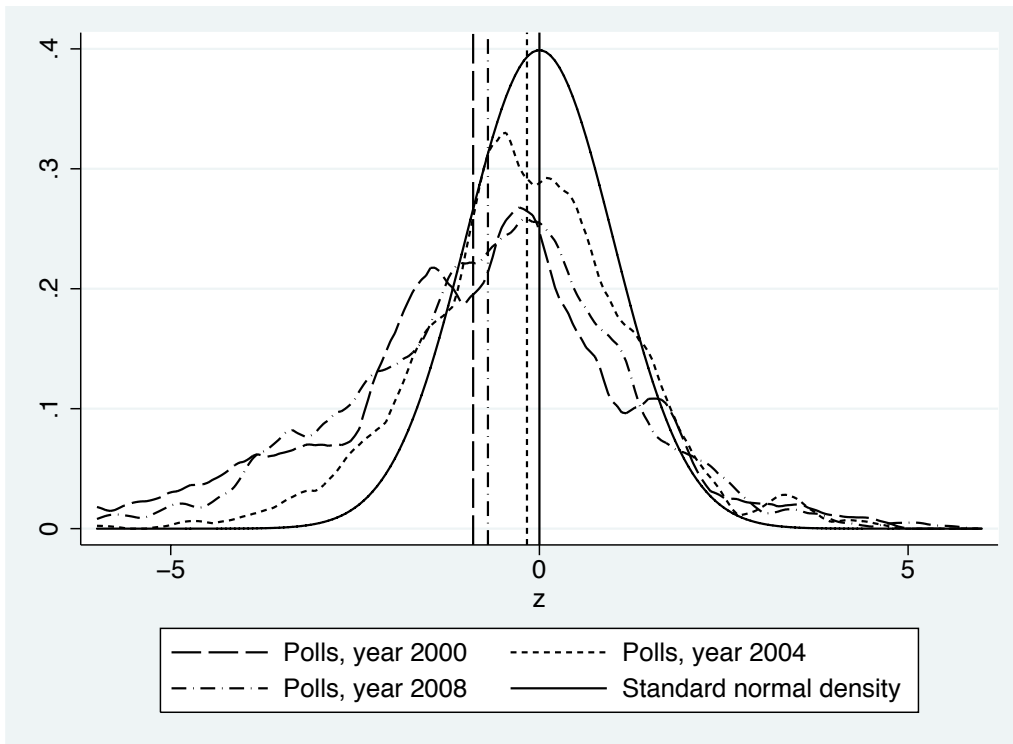
Web Appendix Figure 11: Density Estimates of Standardized Prediction Errors, by Detailed Poll Subgroup

Figure 11 is an extension of Figure 2 with two additional subgroups: polls that sum to 100-102 percentage points, and polls that do not allow third party candidates as an option for respondents. The polls that sum to 100-102 do not look much better than the density of all polls. The polls that exclude third parties show about the same amount of dispersion as polls more generally, but in the case of the Republican share the density is shifted to the right, indicating bias in the direction of over-prediction. The bandwidth for density estimation was chosen by ocular inspection and is 0.2.

Web Appendix Figure 1: Density Estimates of Standardized Prediction Errors, by Election Year



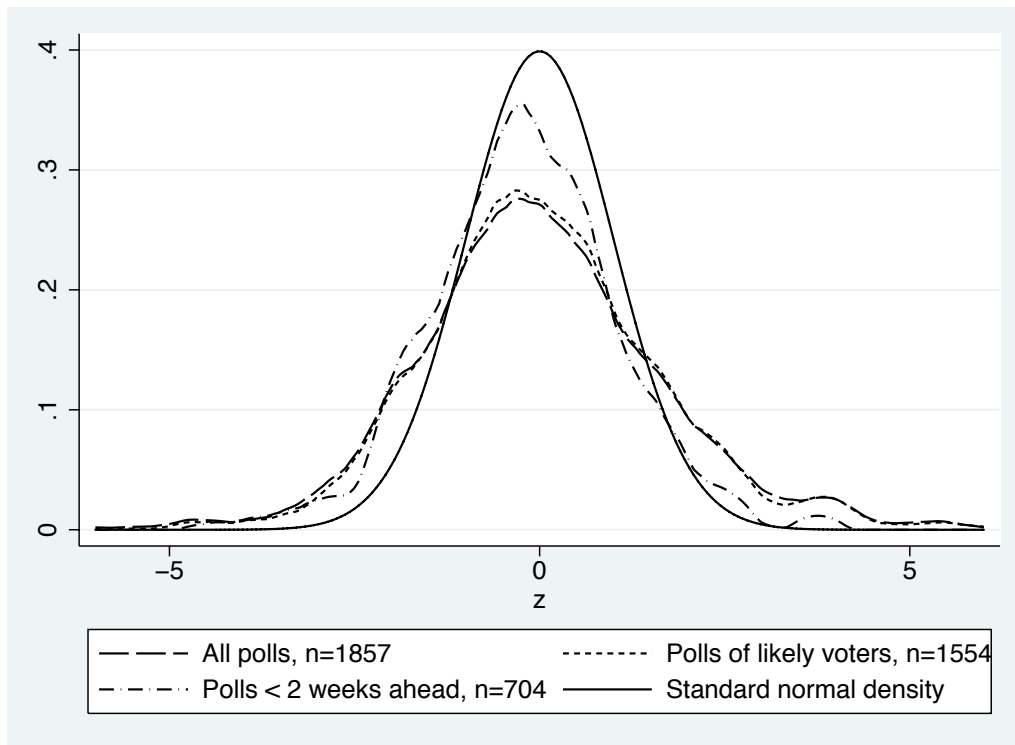
(a) Republican Prediction Error



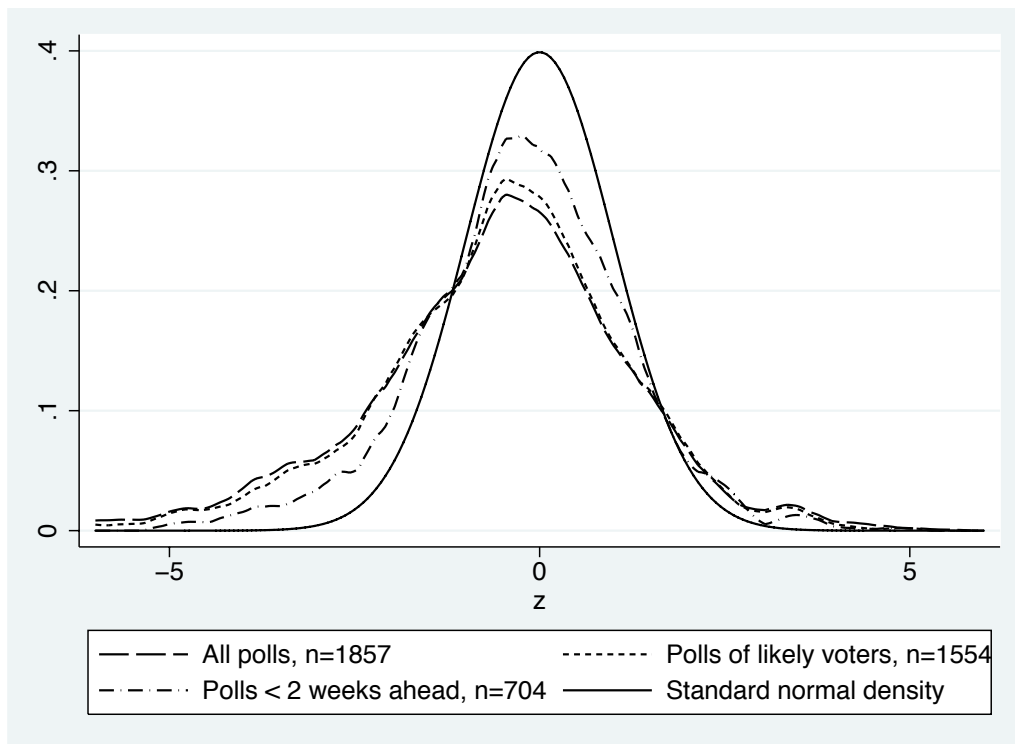
(b) Democratic Prediction Error

The figure displays a kernel density of the standardized prediction errors for presidential state races by election year. The vertical lines are the estimated mean associated with the appropriate density.

Web Appendix Figure 2: Density Estimates of Standardized Prediction Errors, by Poll Subgroup



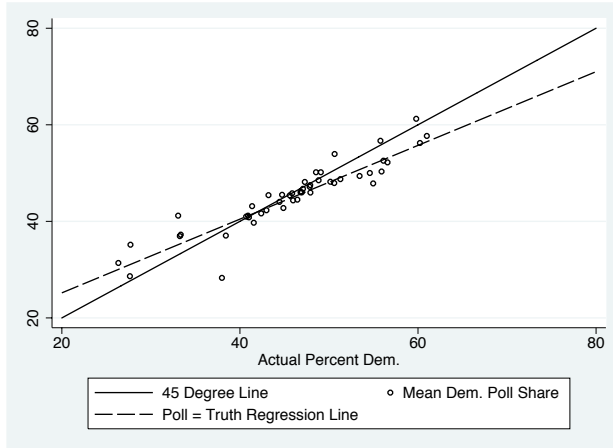
(a) Republican Prediction Error



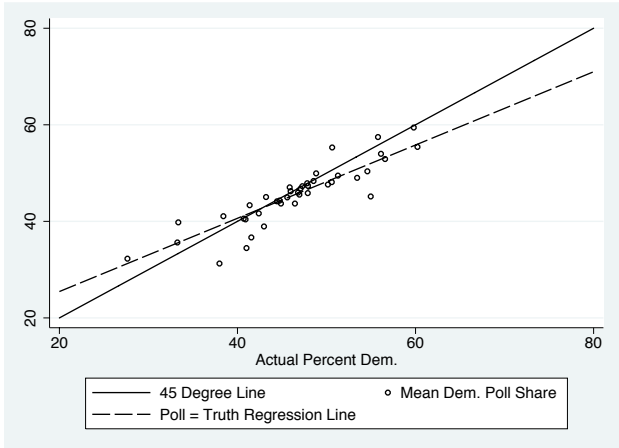
(b) Democratic Prediction Error

The figure displays a kernel density of the standardized prediction errors for presidential state races by poll subgroup.

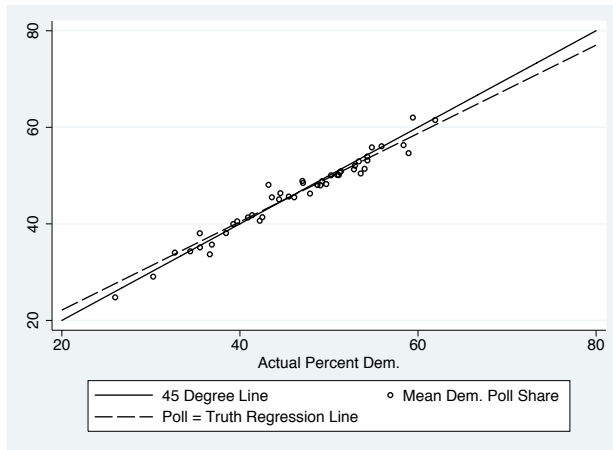
Web Appendix Figure 3: The Relation Between Forecasts and Election Results, Democratic Vote Share



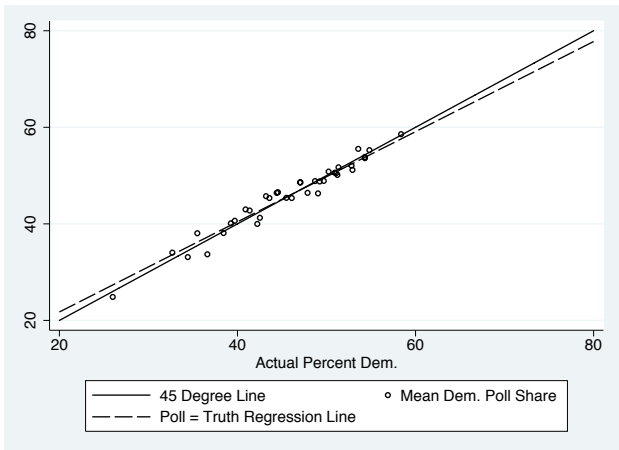
(a) All Polls, 2000



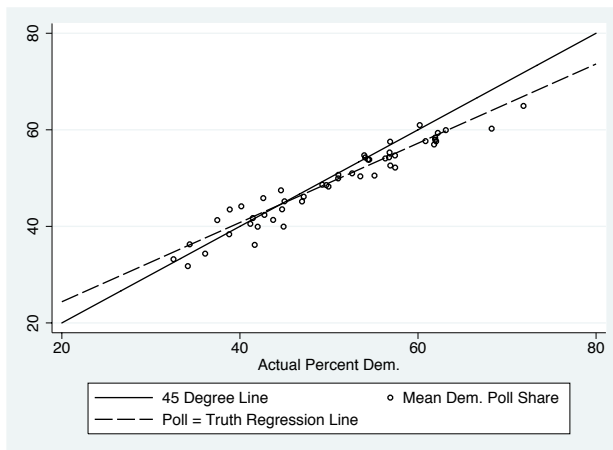
(b) Polls Within Two Weeks of Election, 2000



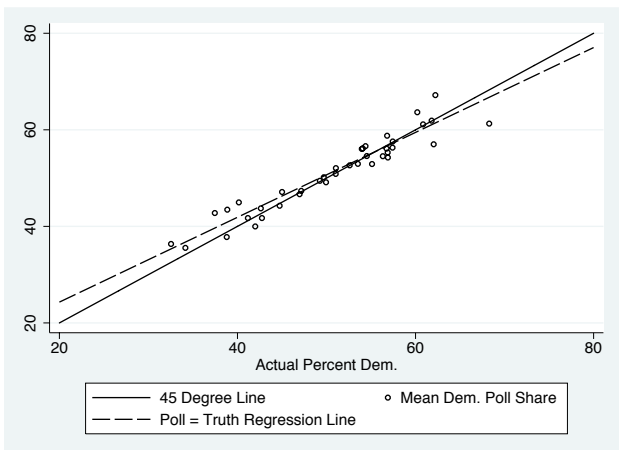
(c) All Polls, 2004



(d) Polls Within Two Weeks of Election, 2004



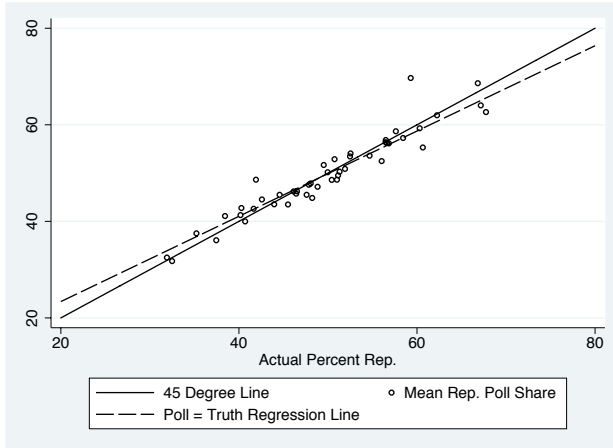
(e) All Polls, 2008



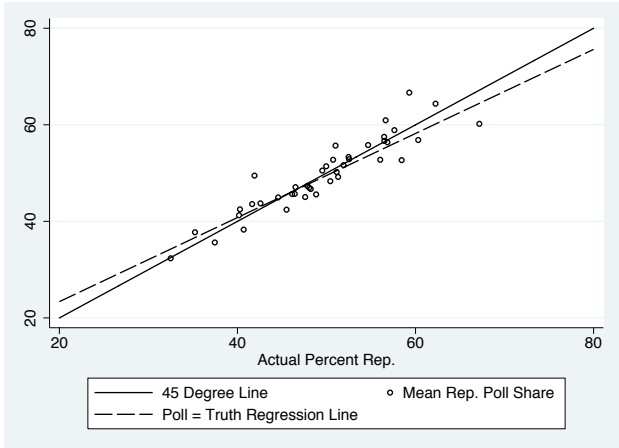
(f) Polls Within Two Weeks of Election, 2008

Each circle represents the mean of all poll results in a statewide election. The dashed line is the estimated line from a regression of the poll prediction on the actual election outcome. The solid line is the 45-degree line.

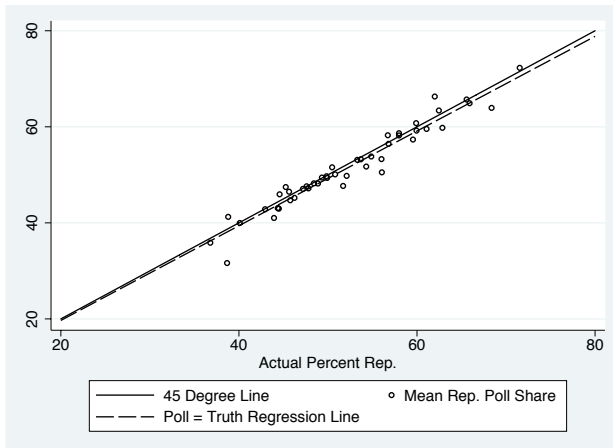
Web Appendix Figure 4: The Relation Between Forecasts and Election Results, Republican Vote Share



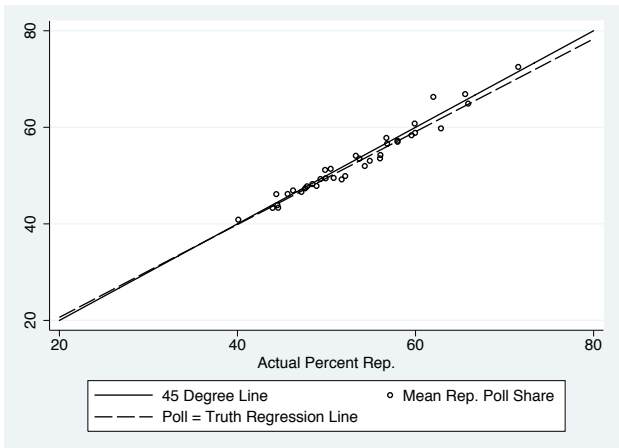
(a) All Polls, 2000



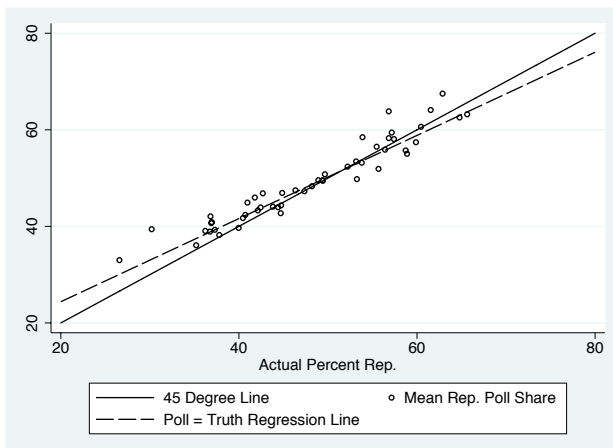
(b) Polls Within Two Weeks of Election, 2000



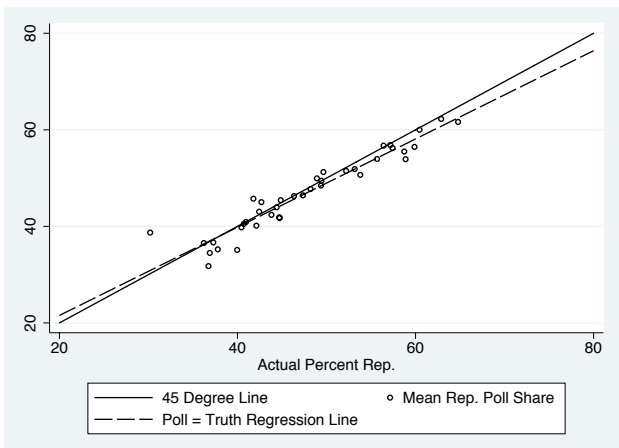
(c) All Polls, 2004



(d) Polls Within Two Weeks of Election, 2004



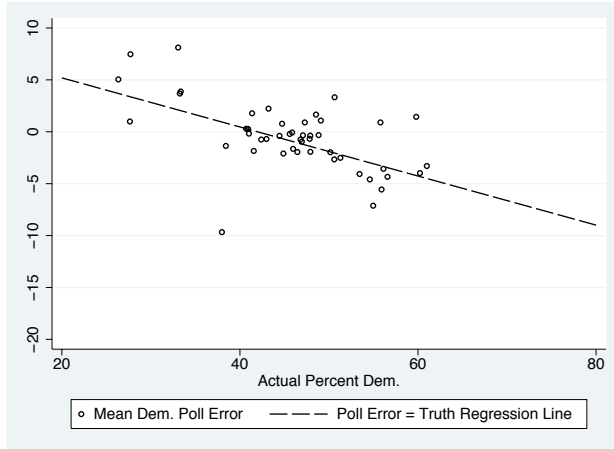
(e) All Polls, 2008



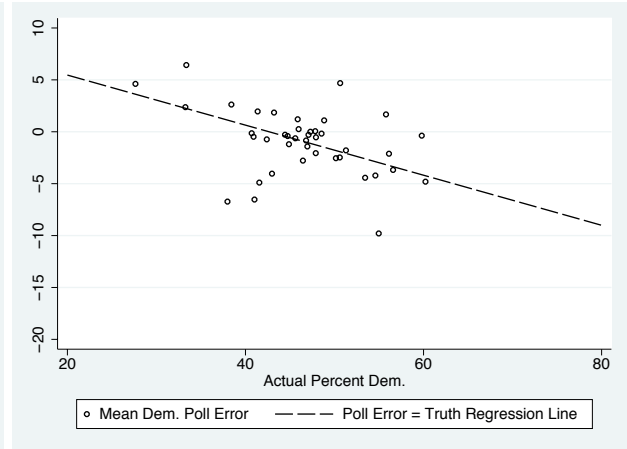
(f) Polls Within Two Weeks of Election, 2008

Each circle represents the mean of all poll results in a statewide election. The dashed line is the estimated line from a regression of the poll prediction on the actual election outcome. The solid line is the 45-degree line.

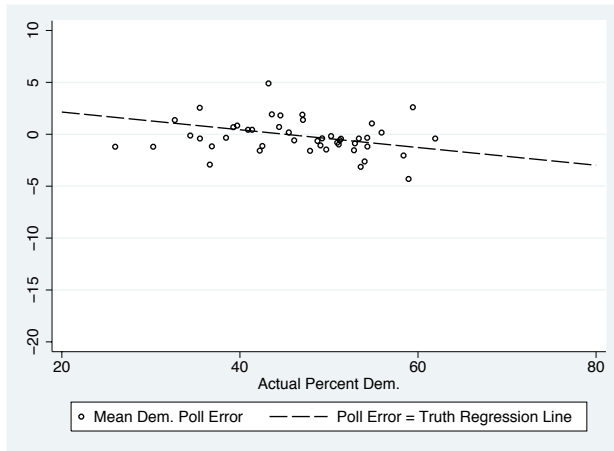
Web Appendix Figure 5: The Relation Between Forecast Errors and Election Results, Democratic Vote Share



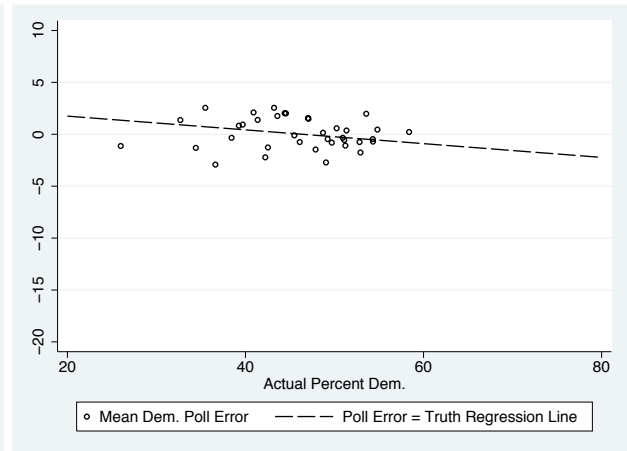
(a) All Polls, 2000



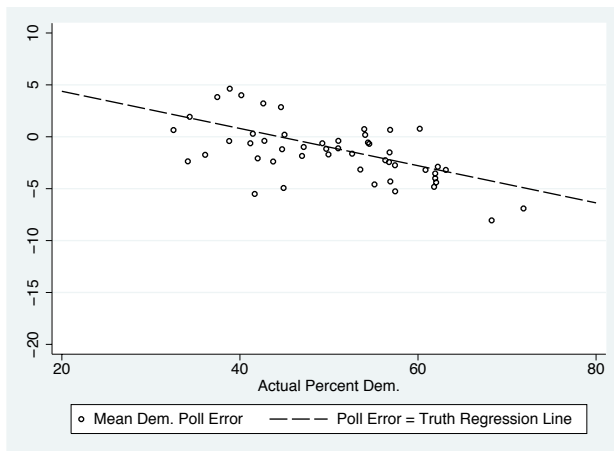
(b) Polls Within Two Weeks of Election, 2000



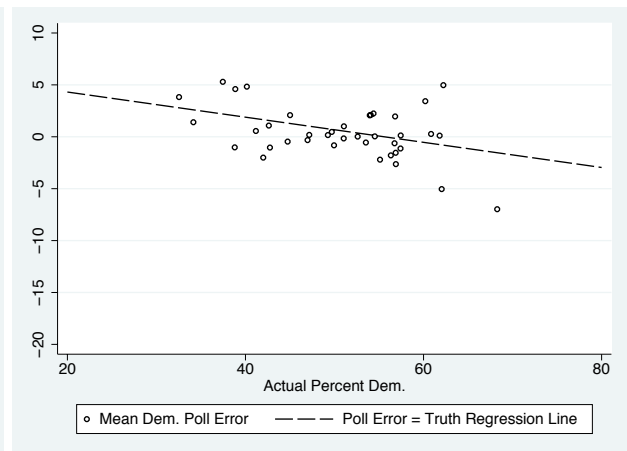
(c) All Polls, 2004



(d) Polls Within Two Weeks of Election, 2004



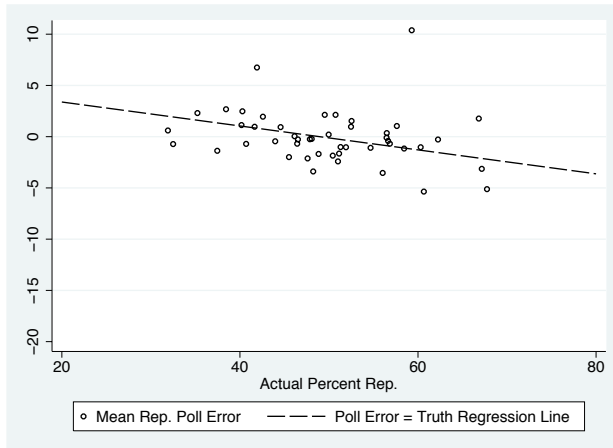
(e) All Polls, 2008



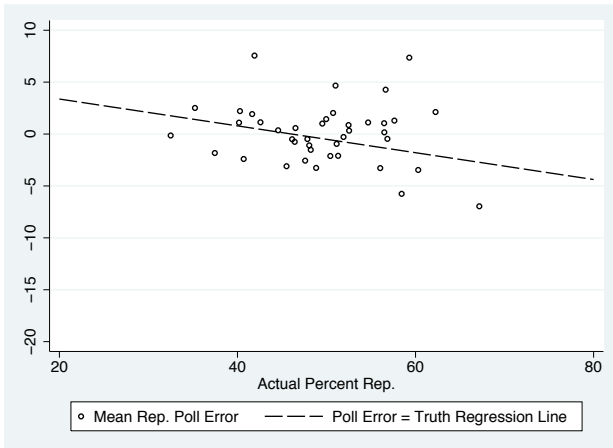
(f) Polls Within Two Weeks of Election, 2008

Each circle represents the mean of all poll prediction errors in a statewide election. The dashed line is the estimated line from a regression of the poll prediction errors on the actual election outcome.

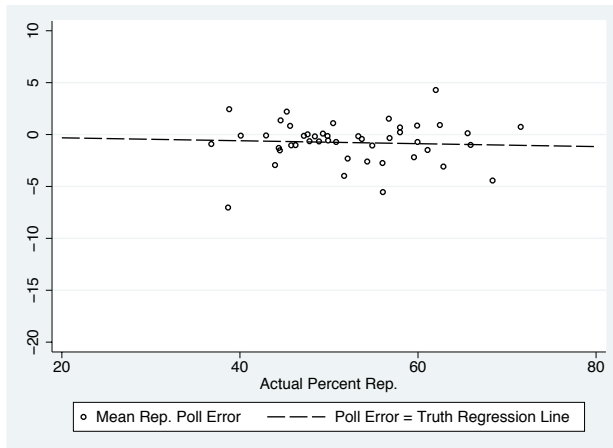
Web Appendix Figure 6: The Relation Between Forecast Errors and Election Results, Republican Vote Share



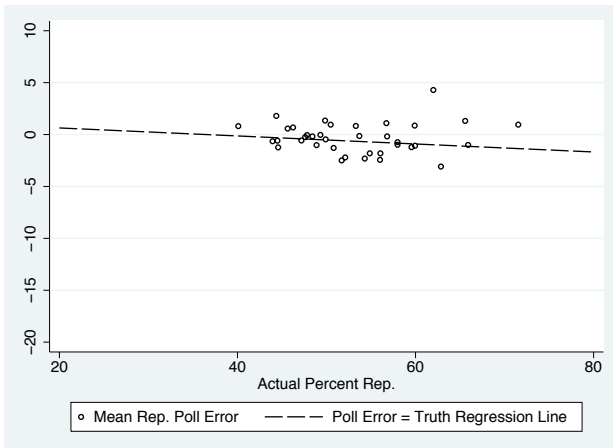
(a) All Polls, 2000



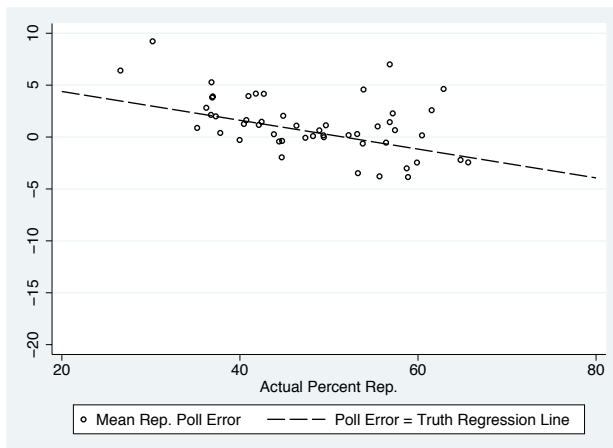
(b) Polls Within Two Weeks of Election, 2000



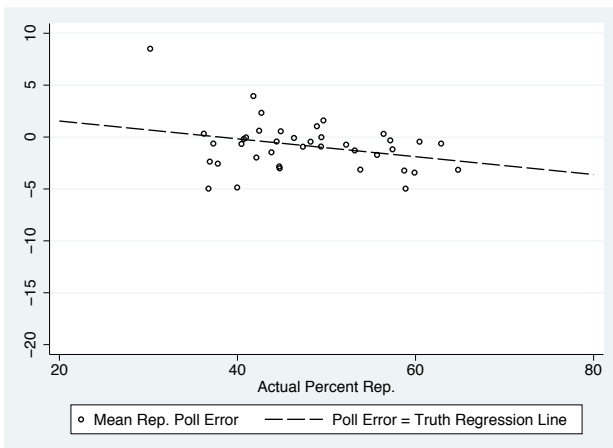
(c) All Polls, 2004



(d) Polls Within Two Weeks of Election, 2004



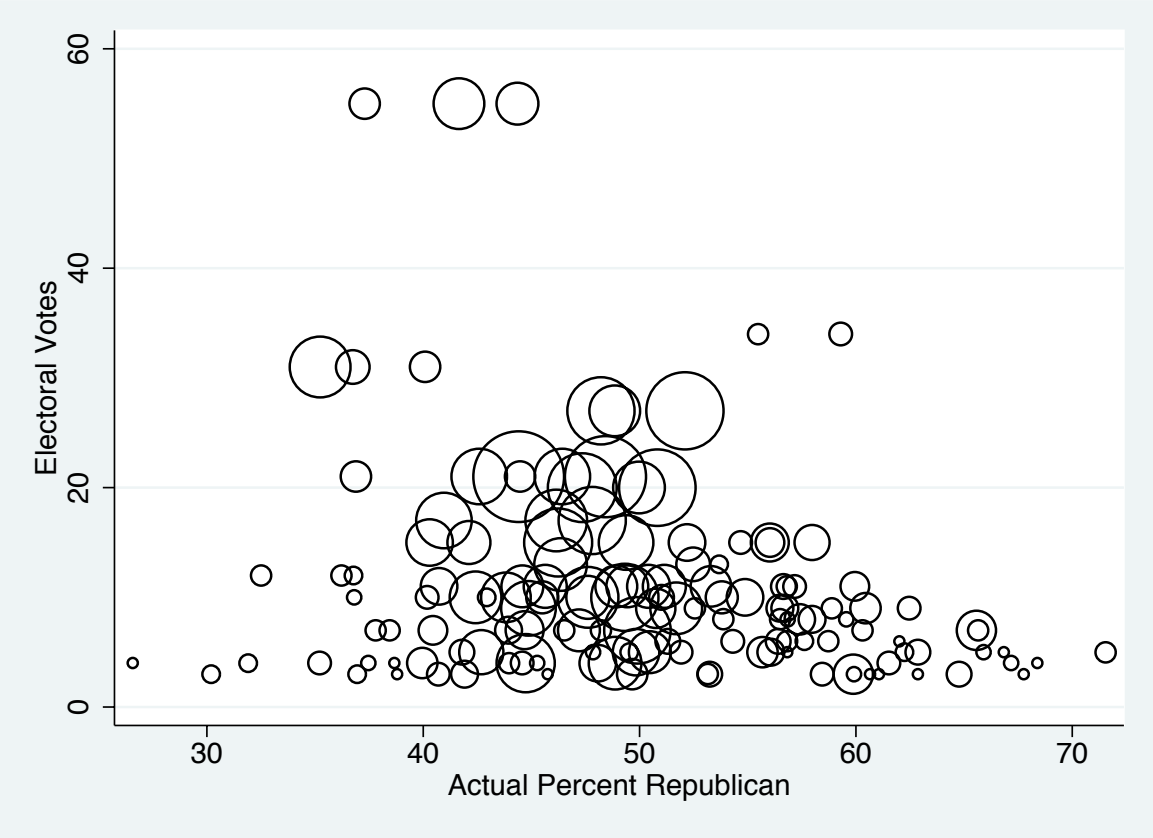
(e) All Polls, 2008



(f) Polls Within Two Weeks of Election, 2008

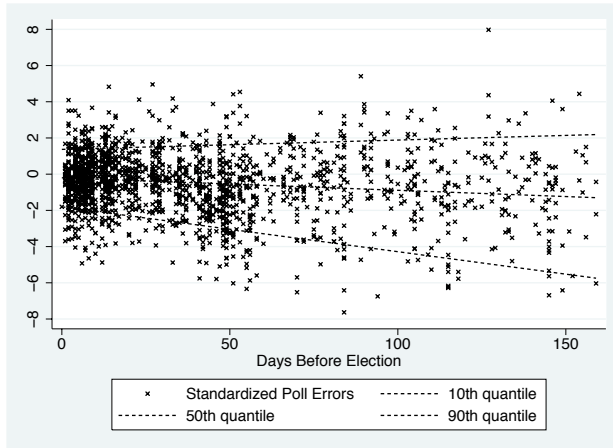
Each circle represents the mean of all poll prediction errors in a statewide election. The dashed line is the estimated line from a regression of the poll prediction errors on the actual election outcome.

Web Appendix Figure 7: Distribution of Polls Across States By Election Result and Number of Electoral Votes

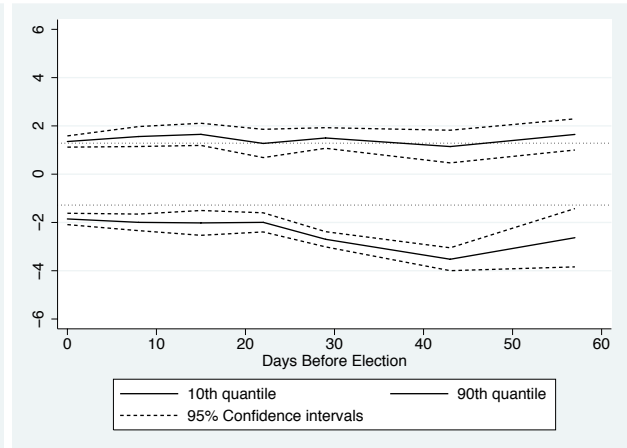


Each circle represents a statewide election. The area of the circle is proportional to the number of polls in that race.

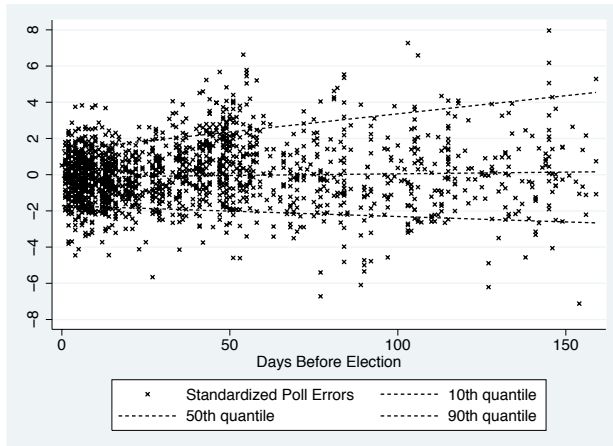
Web Appendix Figure 8: Standardized Prediction Errors Over Time



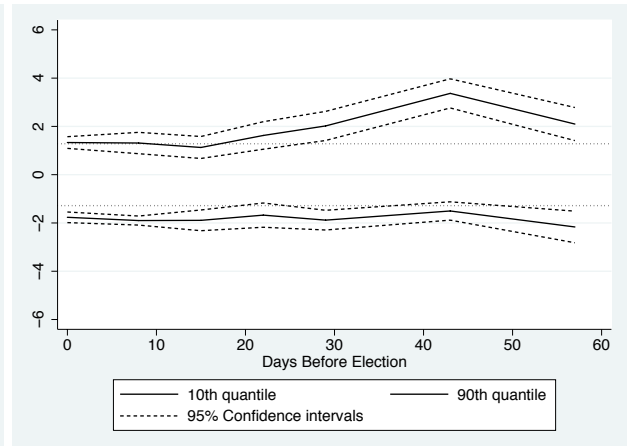
(a) All Democratic Standardized Prediction Errors



(b) All Democratic Standardized Prediction Errors



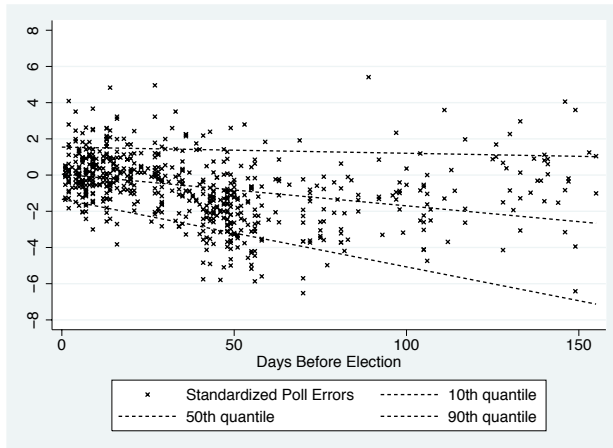
(c) All Republican Standardized Prediction Errors



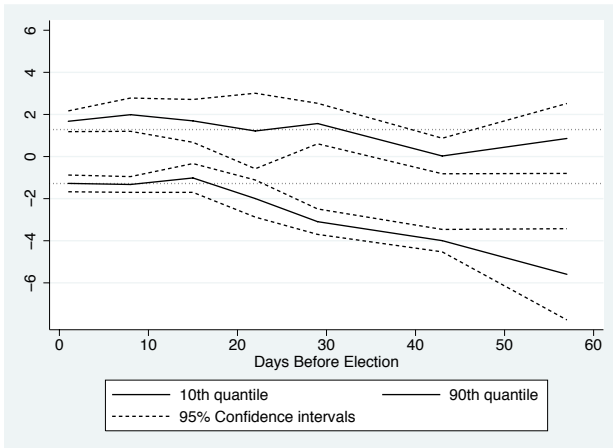
(d) All Republican Standardized Prediction Errors

The figure displays scatter plots of standardized prediction errors for presidential statewide races and quantile regressions at the 10th, 50th, and 90th quantile. The lines in panels (a) and (c) present the results of a quantile regression of the prediction errors on the number of days before the election and a constant term. Panels (b) and (d) present the 10th, and 90th quantiles and associated confidence intervals from a design-adaptive bandwidth quantile regression, limiting the sample to only those polls within 10 weeks of the election. The two dotted horizontal lines in each of panels (b) and (d) indicate the theoretical prediction of the 90th and 10th percentiles under standard normality (1.28/-1.28).

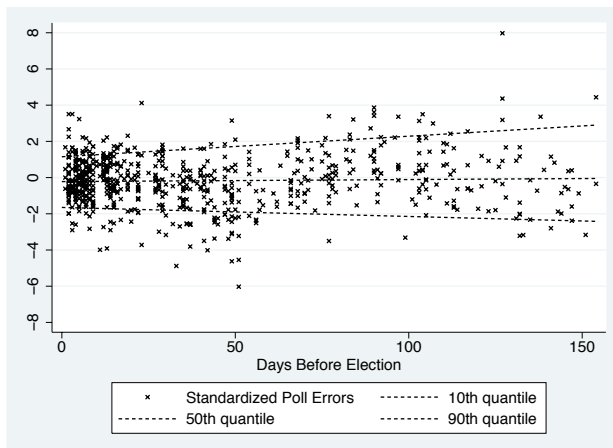
Web Appendix Figure 9: Standardized Democratic Prediction Errors Over Time, by Election Year



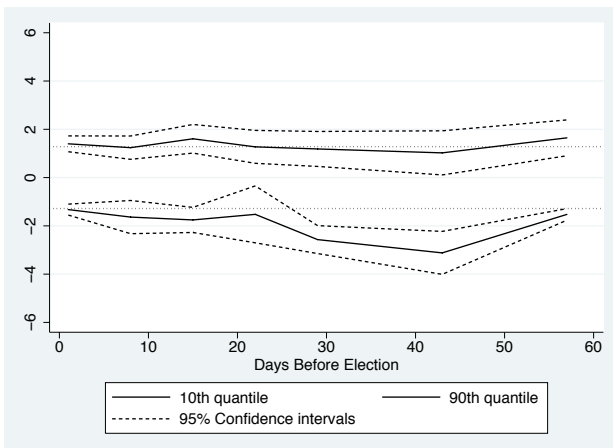
(a) 2008 Standardized Prediction Error



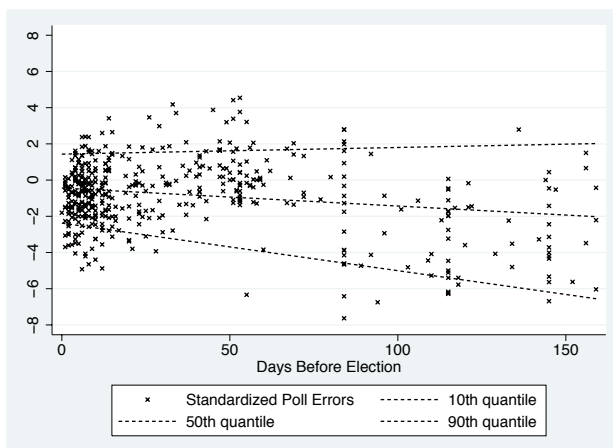
(b) 2008 Standardized Prediction Error



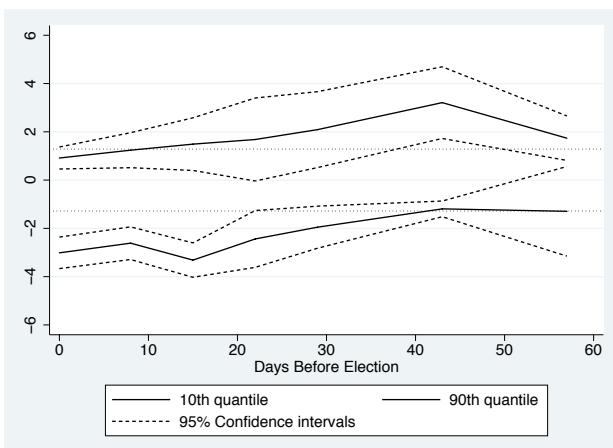
(c) 2004 Standardized Prediction Error



(d) 2004 Standardized Prediction Error



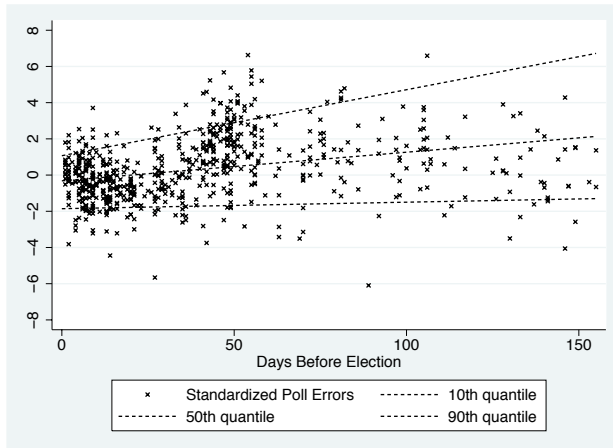
(e) 2000 Standardized Prediction Error



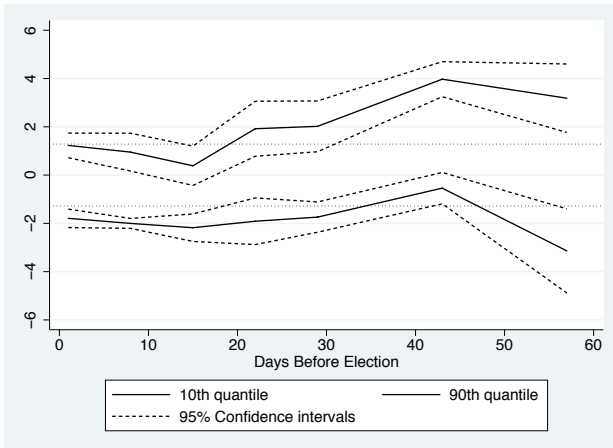
(f) 2000 Standardized Prediction Error

The figures display scatter plots of standardized prediction errors for presidential statewide races and quantile regressions at the 10th, 50th, and 90th quantile for polls separately by election year. The lines in panels (a), (c), and (e) present the results of a quantile regression of the prediction errors on the number of days before the election and a constant term. Panels (b), (d), and (f) present the 10th, and 90th quantiles and associated confidence intervals from a design-adaptive bandwidth quantile regression, limiting the sample to only those polls within 10 weeks of the election. The two dotted horizontal lines in each of panels (b), (d), and (f) indicate the theoretical prediction of the 90th and 10th percentiles under standard normality (1.28/-1.28).

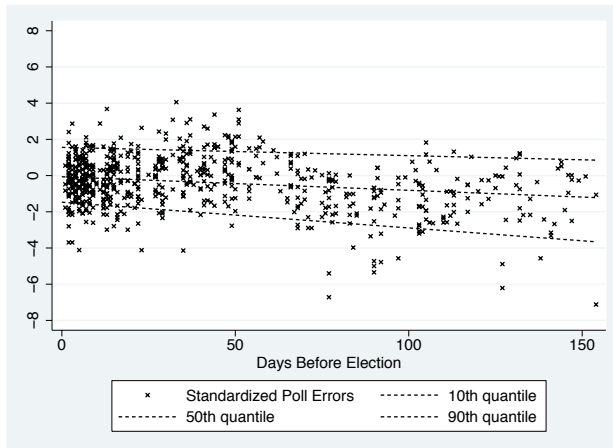
Web Appendix Figure 10: Standardized Republican Prediction Errors Over Time, by Election Year



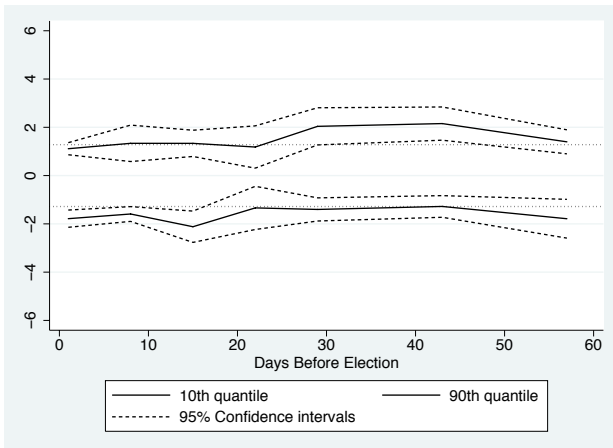
(a) 2008 Standardized Prediction Error



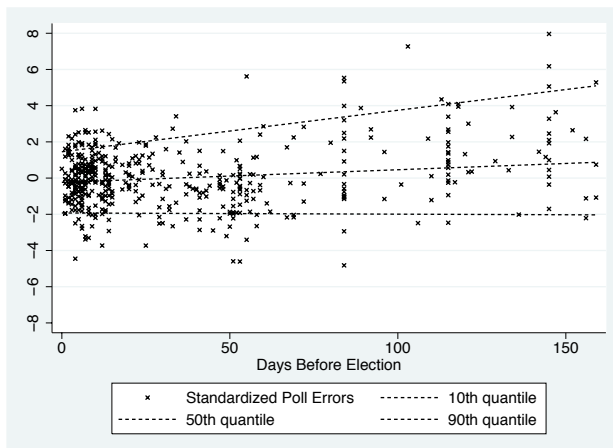
(b) 2008 Standardized Prediction Error



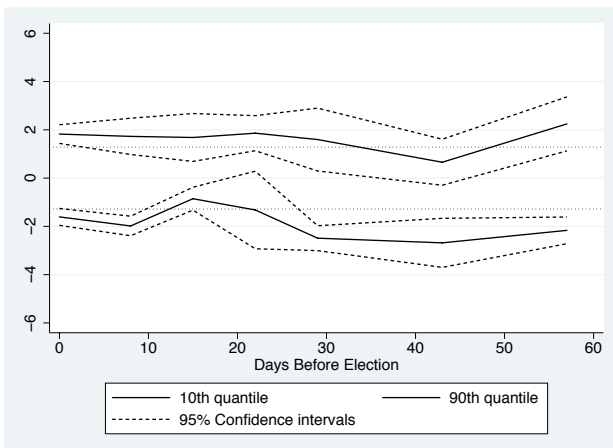
(c) 2004 Standardized Prediction Error



(d) 2004 Standardized Prediction Error



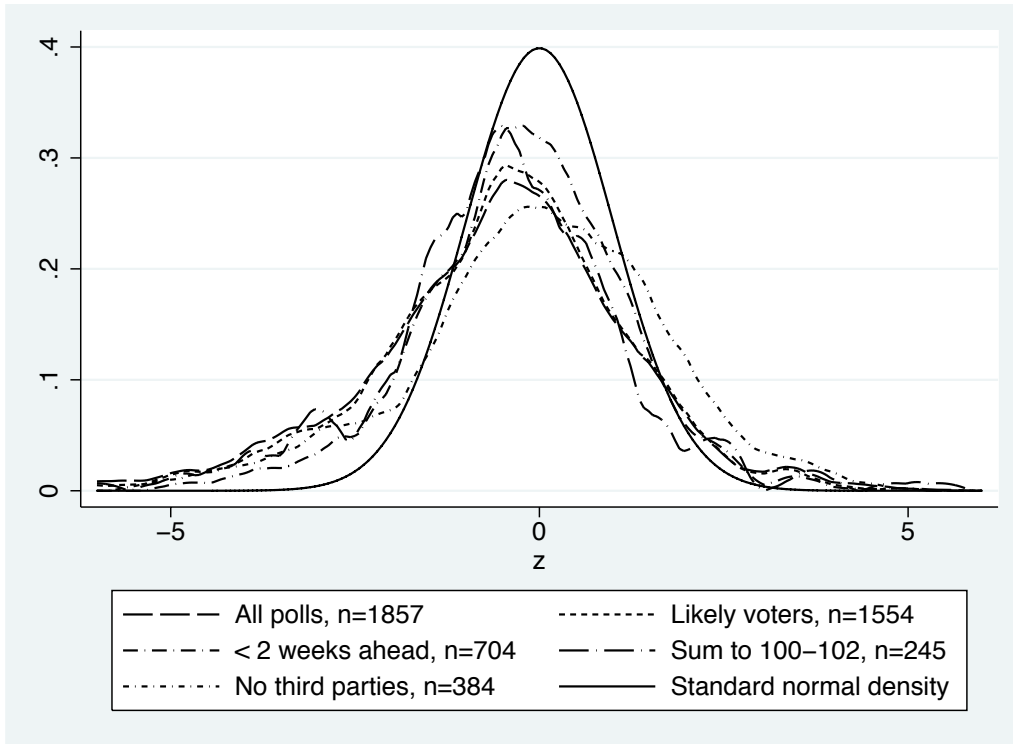
(e) 2000 Standardized Prediction Error



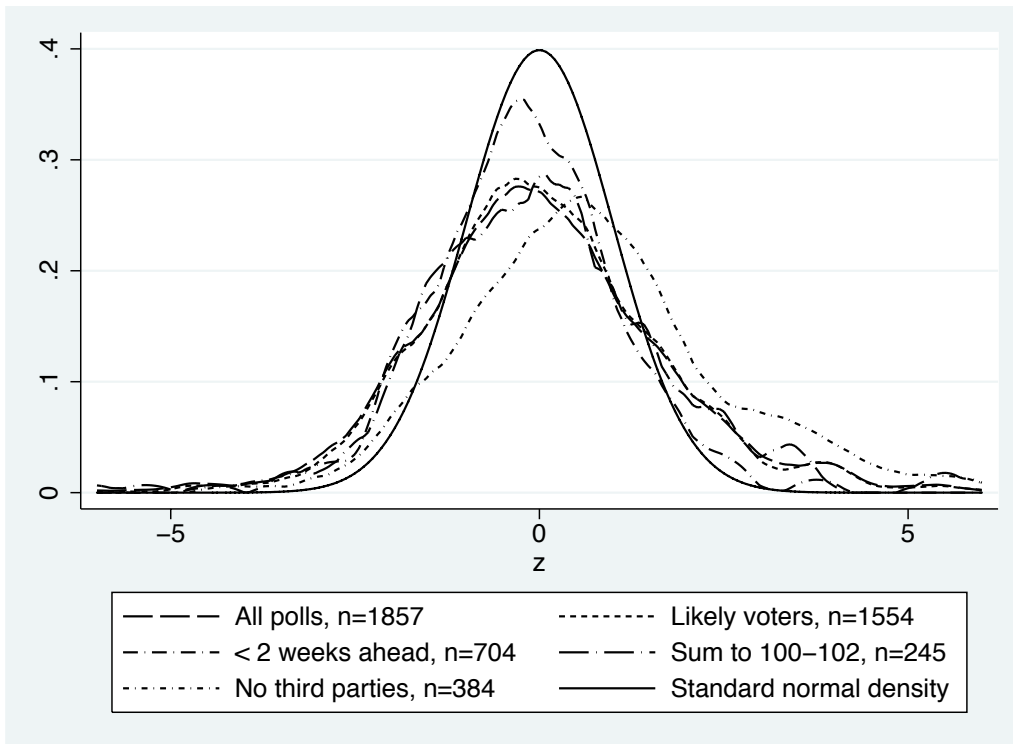
(f) 2000 Standardized Prediction Error

The figures display scatter plots of standardized prediction errors for presidential statewide races and quantile regressions at the 10th, 50th, and 90th quantile for polls separately by election year. The lines in panels (a), (c), and (e) present the results of a quantile regression of the prediction errors on the number of days before the election and a constant term. Panels (b), (d), and (f) present the 10th, and 90th quantiles and associated confidence intervals from a design-adaptive bandwidth quantile regression, limiting the sample to only those polls within 10 weeks of the election. The two dotted horizontal lines in each of panels (b), (d), and (f) indicate the theoretical prediction of the 90th and 10th percentiles under standard normality (1.28/-1.28).

Web Appendix Figure 11: Density Estimates of Standardized Prediction Errors, by Detailed Poll Subgroup



(a) Democratic Prediction Error



(b) Republican Prediction Error

The figure displays a kernel density of the standardized prediction errors for presidential state races by poll subgroup.

Web Appendix: Discussion of Intentions and Polling

Elias Walsh, Sarah Dolfin and John DiNardo

January 30, 2009

I. Probabilistic Intentions

While a large literature (see Crespi (1988) for a nice summary) suggests that “horse race” polls – those that ask respondents about who they intend to vote for in an election – should, if conducted properly and under the right conditions, reflect actual outcomes, an old statistical literature, most recently Manski (1990) suggests the opposite. Manski (1990) observes that if a potential voter is uncertain about who s/he will vote then a simple “intention” question: “who are you likely to vote for” will be biased in general for the outcome even if agents are perfectly rational, etc. The only hope for generating an unbiased prediction of an outcome from intentions data requires asking the question in such a way that allows the voter to express his or her uncertainty.

Instead of asking: *If the election were held today, would you:*

- Vote for John Kerry, the Democratic nominee for president.
- Vote for George Bush, the Republican nominee for president.
- Vote for another candidate.

one should ask the question in terms of *probabilities* for voting for each of the candidates.

It seems worthwhile to ask whether this “theoretical” source of bias can explain much of the bias we observe in actual polls. In a sense, we would like to see the extent to which this purely “statistical” problem addresses the question posed by Gelman and King (1993) – are polls variable only because the questions are posed as intentions instead of probabilities? The purpose of this section of the paper is to investigate the importance of this question by a comparison of responses to “horse race” questions asked the usual way, and the way suggested by Manski’s analysis. Both trends and the reliability of the implied forecast may be quite different for the two sets of questions and this might yield insights as to why polls tend to be biased forecasts of the outcomes.

While this source of bias has been studied extensively for continuous outcomes such as income (see Dominintz and Manski (1997) for a review and example) to the best of our knowledge has not been studied in this context. This problem arises routinely in data of interest to political scientists, economists, sociologists and others and may have implications for broader issues than merely horse race election polling *per se*.

Although “horse race” polls are routinely used to forecast the likelihood that some candidate will win an election, it is well understood in the statistics literature that even in the “best case” there is no reason to suppose that “intentions” (“I am likely to vote for candidate X”) should yield unbiased forecasts of actual behavior. (Manski, 1990)

We first focus on a “best case” scenario and illustrate with some simple numerical examples why

1. Polls should be biased in general.
2. Even large positive changes in poll results over time do not necessarily indicate increased support for the candidate.

In doing so, we focus only on the possibility that some individuals are uncertain about who they will vote for. We assume that all the other possible problems (sample selection biases, question ordering, etc.) that have been cited in the literature are solved.¹ As a rule, assuming something worse than the “best case” results in an even greater bias and for reasons of brevity and clarity we omit that discussion here.

A The Best Case

Following Manski (1990), let i be a binary indicator denoting an intention – “talking about the presidential elections in November, for whom are you likely to vote – George Bush?” and let y be the indicator corresponding to the actual behavior (the individual votes for Bush). Letting s denote the information available at the time of the survey to the respondent and let z denote the events that have not yet occurred but that will affect his future action.² Let $P_z|s$ denote the objective distribution of z conditional on s . Let $P(y|s)$ denote the objective distribution of y conditional on s . The event $y = 1$ occurs \iff the *realization* of z is such that $y(s, z) = 1$.

In the *best* case, we assume rational expectations: this means the respondent *knows* how they will act depending on the possible realizations of z and that they also know $P_z|s$ – that is they know the stochastic process generating z – in words, the respondent knows the correct distribution of the behavior influencing events z and moreover uses that information optimally. To take a concrete example, suppose z is the public exposure of a scandal involving “morals” or sexual behavior of a candidate. This assumption is the requirement that I know how I would behave if my candidate were involved in a scandal and the the probability that I would learn about such a scandal before election day.

The second aspect of the “best case” scenario is that the respondent states her best point prediction of her behavior. The best prediction depends on her “loss function” associated with either $(i = 1, y = 0)$ and $(i = 0, y = 1)$. Manski observes that under these two sets of assumptions the responses satisfy:

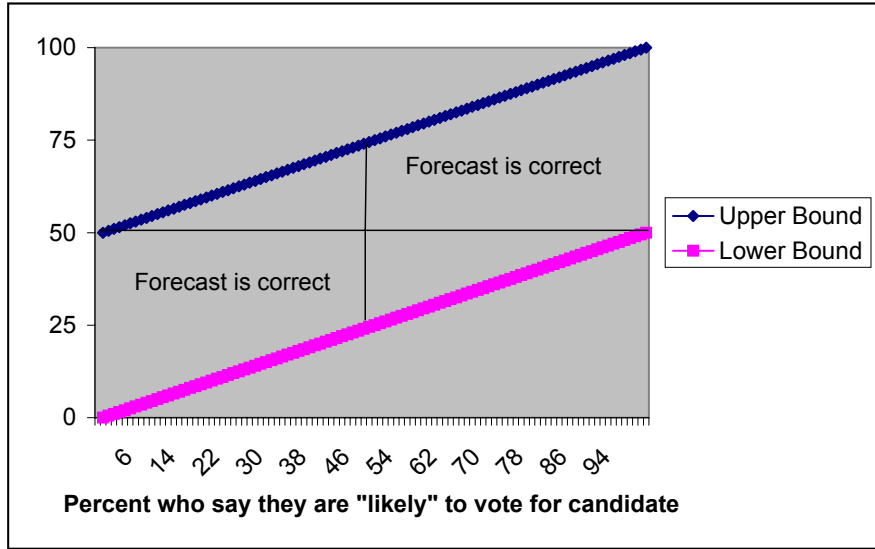
$$\begin{aligned} i = 1 &\implies P(y = 1|s) \geq \pi \\ i = 0 &\implies P(y = 1|s) \leq \pi \end{aligned} \tag{1}$$

In words, if the action y is “voting for candidate X”, then a respondent tells the interviewer that she will vote for candidate X if the probability that she will do so is greater than π . If both possible errors are equally “costly” than $\pi = .5$ Specializing to the case of horse race polls, the object of the poll is to learn the probability $P(y = 1|i, s)$. As Manski observes, however, the pollster’s data on “intentions” does not identify that probability. Even in this “best case” – assuming that persons have identical loss functions – they only imply a “bound”. As Manski shows:

¹See for example, Gelman and King (1993) or Ottaviani and Norman (2006) for discussions.

²To make the problem even more simple, we assume that a person’s participation is known with certainty. Allowing for uncertainty in participation only strengthens the negative result.

Figure 1: The Bounds Implied by “Intentions” are not tight: A comparison of intentions with outcomes



$$P(y = 1|s, i = 0) \leq \pi \leq P(y = 1|s, i = 1)$$

expresses *all* the information in intentions data.

Figure 1 displays the bounds implied by the data assuming no sampling error, that individuals have identical symmetric loss functions, and that there is no “new information” s between the time the poll is taken and the behavior occurs. The dependent variable is the actual voting outcome on election day.

The lower right and upper left triangles that lie within the polygon formed by the bounds indicate that 25 percent of the area within the bounds fail even to cover the correct binary prediction of the outcome. Note that it would be incorrect to draw the inference that the polls would get it right 75 percent of the time in this best case. Rather, the correct inference is that the correct bounds do not have to cover the correct binary prediction of the election outcome. Of course, if the sample is not a random sample, new information occurs between the poll and the event, or that there is a double uncertainty (i.e. the voter does not know for certain whether s/he votes) the bounds could easily be *completely* uninformative.

B A Rise in the Polls Doesn’t Necessarily Imply Increased Support

Observe that we have gone a bit beyond even the “best case” in this simple illustration. As Manski observes (and as was observed earlier by Juster (1966), for example) it has been well known in the statistical literature that such polls will *not* be unbiased in general, even in this *best case*. As a consequence, a poll is especially unsuited to assessing “trends” in voter support for a candidate, even when the electorate is composed of Bayesian statisticians with correct rational expectations.

The following contrived example, although not altogether unreasonable, shows an example where support for candidate X is falling (measured as what would actually have happened if an election had been conducted), at the same time the polls are showing a massive increase in support for the candidate. For simplicity, we have three types of voters. Type “C” voters strongly support candidate X, type “B” voters less strongly support candidate X, and type “A” voters strongly oppose candidate X. Between the two periods, type “A” voters grow much more strongly opposed to candidate X, and type “B” voters slightly shift in favor of candidate X. As a consequence, the polling shows a large increase in support for candidate X from period 1 to period 2, even as the actual probability of X being elected fell over this time!

Table 1: Polls show increased support, when support is falling

Voter “Type”	Fraction in Population	Time Period 1		Time Period 2	
		Probability vote for X	Response to Pollster	Probability vote for X	Response to Pollster
A	0.25	0.4	0	.1	0
B	0.5	0.46	0	.51	1
C	0.25	0.8	1	.8	1
		Actual Outcome	0.53	Actual Outcome	0.48
		Poll Result	0.25	Poll Result	0.75

C Voter Participation

The above analysis applies *mutatis mutandis* to an analysis of voter participation as a separate inquiry. As far as we have been able to ascertain, polling organizations routinely use a binary measure of whether or not an individual is likely to vote. Again, if the decision to participate is uncertain, in general there is no reason to believe that restricting to the sample to “likely voters” or “registered voters” (the two most frequently used screens in practice) will yield an unbiased rate of participation.

Moreover, since – in the simplest model – the act of the voting for a specific candidate is the product of two uncertain decisions (a decision to support the candidate, and the act of going to the polling booth) it is clear that treating the corresponding sets of intentions as certain – i.e. binary – is biased as a forecast of the actual vote or the “strength” of the support for a candidate.

References

- Crespi, Irving**, *Pre-Election Polling: Sources of Accuracy and Error*, New York: Russell Sage Foundation, 1988.
- Dominintz, Jeff and Charles F. Manski**, “Using Expectations Data to Study Subjective Income Expectations,” *Journal of the American Statistical Association*, September 1997, *92* (439), 855–867.
- Gelman, Andrew and Gary King**, “Why are American Presidential Election Campaign Polls So Variable When Votes are So Predictable?,” *British Journal of Political Science*, October 1993, *23* (4), 409–451.
- Juster, Thomas**, “Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design,” *Journal of the American Statistical Association*, 1966, *61*, 658–696.
- Manski, Charles F.**, “The Use of Intentions Data to Predict Behavior: A Best-Case Analysis,” *Journal of the American Statistical Association*, December 1990, *85* (412), 934–940.
- Ottaviani, Marco and Peter Norman**, “The Strategy of Professional Forecasting,” *Journal of Financial Economics*, August 2006, *81* (2), 441–446.

Web Appendix: Poll Questions

Elias Walsh, Sarah Dolfin and John DiNardo

January 30, 2009

Our study design consisted of the following two pairs of questions:

For the first demographically balanced half-sample, the questions are designed to allow respondents to express probabilistic intentions:

1. *Are you a registered voter? If yes:*

- *Given your other obligations, on a scale of 0 to 100 what is the chance that you will actually cast a vote for president? If you are certain you will vote, state "100." If you are certain you will not vote, state "0". If there is a 40 in 100 chance you will vote, state 40, and so on.*

If no,

- *Given your other obligations, what is the chance that you will register to vote **and** vote for president in November 2004. Use a scale of 0 to 100. If you are certain you will register and you will vote, state "100." If you are certain you will not register, or you will register and not vote, state "0". is a 40 in 100 chance you will both register and vote, state 40, and so on.*

2. *Regardless of whether or not you are likely to vote in the presidential election, given what is likely to happen during the course of the campaign, on a scale of 0 to 100 what is the likelihood that you would vote for John Kerry, George Bush, or some other candidate for president?*

The sum of your answers should be 100. For instance, if there is a 40% chance you would vote for John Kerry and a 40% chance you would vote for George Bush, and a 20% chance you would vote for someone else, your response should be:

John Kerry	40
George Bush	40
Other Candidate	20

If you are certain that you would vote for Ralph Nader (or a candidate other than Bush or Kerry), your response should be:

John Kerry	0
George Bush	0
Other Candidate	100

For the other demographically balanced half-sample, the two questions are designed to mimic typical poll practice.

1. *Are you registered to vote?*

If yes:

- *Are you likely to cast a vote for a presidential candidate in the 2004 election?*

If no,

- Are you likely to register in time for the election **and** cast a vote for a presidential candidate in the 2004 election?

2. Regardless of whether or not you are likely to vote in the presidential election, and given what is likely to happen during the course of the campaign, for whom would you vote:

- Vote for John Kerry, the Democratic nominee for president.
- Vote for George Bush, the Republican nominee for president.
- Vote for another candidate.