Reliable Compared to What?

Empirical Tests of the Accuracy of Election Polls, 2002

Joel David Bloom

Oregon Survey Research Laboratory
441 McKenzie Hall
University of Oregon
Eugene, OR 97403-5245

Political Science Department
936 Prince Lucien Cambpell
University of Oregon
Eugene, OR 97403-1284

jbloom@uoregon.edu
(541) 346-0891

http://www.uoregon.edu/~jbloom

(DRAFT REVISION – June 20, 2003)

Abstract: I analyze a data set of 232 polls from Senate elections in 15 states in which the winning margin was less than 20% in 2002. By one measure (net partisan bias) the polls seem to be reliable, while by others – most notably median error compared to election results and the percentage of polls that fall outside their own published margin of error compared to the election results – they performed quite poorly. While adjusted reported margins of error averaged 4.4% in my sample, I estimate observed margins of error of 6.1% and 7.4% based on expected area beneath a normal curve. On the strictly subjective measure of predictive accuracy, these surveys are also a mixed bag. One important, although not surprising, finding is that published partisan polls reflect a partisan bias. The other fairly strong finding here is that despite the criticism to which they were subjected, Zogby International and SurveyUSA performed at roughly the same level as other nonpartisan polling organizations in 2002. By most measures, Zogby did just slightly more poorly than the norm, while SurveyUSA did somewhat better.

This paper has been revised and expanded on since its presentation at the 2003 Annual Meeting of the American Association of Public Opinion Research, May 2003, in Nashville, TN. Comments and suggestions are welcome. I will be happy to share the data set on request in exchange for any corrections or additions you may have.

Introduction: Unreliable Compared to What?

Pollsters were the favorite media punching bag after the November 2002 elections, in which Republicans won somewhat more Senate races and a good deal more gubernatorial races than most had predicted. This was not the first time that election pollsters had come under such fire. Surprisingly strong Democratic party showings in the 1998 and 2000 congressional elections triggered similarly strident criticisms of the polls and those who conducted them.

Indeed, in the Senate races there appeared to be some major misses by the pollsters. In New Hampshire, most polls showed Democrat Jeanne Shaheen in a very tight race with (and often ahead of) Republican John Sununu who ended up winning by more than 4 points. In Georgia, most polls showed Democratic incumbent Max Cleland ahead, but Republican challenger Saxby Chambliss won by nearly 7 points. In the Colorado rematch between Republican incumbent Wayne Allard and challenger Tom Strickland, polls showed an excruciatingly tight race, but Allard ended up winning by more than 5 points. Finally, in Texas, many polls predicted a very tight race between Republican John Cornyn and Democrat Ron Kirk, with some polls even showing Kirk slightly ahead. In the end, Cornyn ended up winning by nearly 12 points.

Two prominent polling organizations were especially controversial in 2002 – Zogby International, which was strikingly wrong in Colorado and a few other states, and Survey USA, which came under suspicion for its practice of conducting interviews using an automated system rather than live interviewers. In addition, it is often noted but seldom analyzed systematically, that polls conducted by partisan pollsters often seem to lean in the direction of their partisanship. So I will also take up the question of whether polls conducted by these two firms and partisan organizations in particular were on the whole as reliable as those conducted by other organizations.

The appropriate question is not whether some polls were inconsistent with election results; that is to be expected. The question is whether the gap between poll results and election results was larger than we would expect given the laws of statistical probability and the presence of observable changes over time, as well as important differences in polling techniques.

I attempt to answer these questions using a data set of 237 election polls in 15 states in which the eventual margin of victory was less than 20 percentage points (excluding Louisiana, where incumbent Mary Landrieu had a margin of well over 20% but still faced a runoff and Minnesota where the death of incumbent Paul Wellstone less than two weeks before the election makes any comparisons of election results to polls impossible). Only polls from September-through October were included with the exception of New Jersey, for which only polls with Frank Lautenberg as the Democratic nominee were included.

Since this phase of the project is a first look at the data, I do not yet employ particularly sophisticated methodologies; however I plan on examining various more comprehensive options after presenting this preliminary analysis and getting feedback on it.

[A more complete literature review section is forthcoming here.]

Why We Might Expect Election Polls to be Problematic

A number of factors suggest that election polls should be less accurate than polls on other issues. The most important reason for this is that unlike standard opinion polls, election polls attempt to sample from an unknown and unknowable population, those who will vote on election day.

First of all, with only between 36% and 40% of the adult population actually voting in 2002, (McDonald, 2003) election polls are targeting a rare, or at least quasi-rare population. Many individual states had far lower turnout, with Texas' 29% the lowest of any state with a contested state-wide election (both the gubernatorial and Senate races were hotly contested). While voters are not nearly as rare as many population subgroups, they are sufficiently rare as to require survey researchers to employ sampling techniques used for rare populations, such as screening questions or sampling from lists of known population members.

Second, and perhaps more importantly, when pre-election polls are in the field voters are a population that technically *does not yet exist*. While some individuals are nearly certain to vote and others are nearly certain not to (with a great many in between), the population of those who will actually vote is, prior to the deadline for voting, is not yet a population, but *in the process of becoming one.* Thus, the population of voters is not only *rare* or *quasi-rare*, but also *latent*. While a great deal of literature exists on the challenges posed by sampling *rare* populations, no such literature exists on sampling of *latent* populations.[1] Needless to say, attempting to sample from a population that does not yet exist presents a rather unique set of challenges to election pollsters, with an impact on reliability that is both potentially very large and certainly impossible to estimate in advance.

While this paper is the first use of the terms "rare" and "latent" to describe the target population in a pre-election survey, others have noted the problems associated with these features. As Traugott and Lavrakas put it:

> While there is strong scientific and statistical basis for drawing samples and constructing questionnaires, estimating who will vote on Election Day is an area where the practitioner's art comes into play. There is no standard, widely accepted way for estimating a person's likelihood of voting. Most polling organizations combine the answers to several questions to estimate the likely electorate, and some methods work better than others. (2000, p. 14)

Norman Ornstein puts it more colorfully:

> We try to portray polling as a science, but it's a witchcraft kind of art. When it comes to the midterm elections, we're trying to predict how 35% of the electorate will vote, but we don't know which 35% will turn out. It's beyond embarrassing. (Neumann, 2002)

As a result, the largest differences observed among polls may not be due to differences in their polling methods (although these differences are considerable), but rather the methods for determining who is a likely voter, and sometimes the weights applied to various demographic or political subgroups. (See, e.g., Traugott and Lavrakas, 2000 and Crespi, 1988.)

[1] It is challenging to imagine another instance of a latent population that a researcher might attempt to survey. The closest parallel might be a survey of a population consisting of individuals who are, based on a number of characteristics, considered to be at risk of developing a particular disease or syndrome, but even there the parallel is limited since no one would be attempting to determine their intention of developing the disease.

These post-survey manipulations are arguably necessary – after all, we know for a fact that not everyone with an opinion will vote. But even if they are not witchcraft, they are educated guesses and add potentially non-random bias to reported survey results. Unfortunately, since these techniques for determining likely voters are often closely-guarded trade secrets, analysis of differences resulting from these methods is impossible.

The fact that pollsters can't really know in advance of the election who will be in the population of those who will actually vote makes the laws of sampling error theoretically inapplicable. In other words, when we sample from a known population of one million, for example, we can use our standard sampling tables to say that a sample of 400 will be associated with a sampling error of 4.9%, 600 with 4.0%, 1,000 with 3.1%, and so on. And, indeed, election polls routinely report these figures as if they have drawn a sample from a known population.

The fact that they cannot actually do so means that it is quite possible that election polls might on average produce a range of results that appear to be less reliable than sampling error would predict, even if that observed unreliability might still be due only to sampling-*related* error). In such a framework, non-coverage bias and nonresponse bias can also play major roles (this paper will not be able to address the problem of diminishing response rates in election polls but by many reports it has fallen to problematic levels).

At the same time that election polls face all these challenges not faced by standard attitude polling, they also face a much stricter test not faced by other surveys– if your poll shows a certain percentage of the population favoring a tax cut no one will be able to present you with a number that is self-evidently the right answer with which to compare it. In election polls, of course, one faces the prospect of an impending election that presents just such a test.

If, despite these challenges, election polls indeed turn out to be as reliable or more reliable than the standard response error formulae predict, this would be quite a tribute to the skills and talents of election pollsters.

The Data Set

With the help of my research assistant, Christopher Smith-Hill, I compiled a sample of 232 polls from a number of publicly available internet locations, including: "D.C.'s Political Report" (http://www.dcpoliticalreport.com/2002/polls02.htm), *National Journal's* "Poll Track" (http://nationaljournal.com/members/polltrack/2002/races/), Kiva Communication (http://www.kivacom.com/US_Senate_Races_2002.htm), and "Our Campaigns" (http://www.ourcampaigns.com/cgi-bin/r.cgi/SenateList.html?) as well as the data set provided by the National Council on Public Polls (O'Neill, Mitofsky and Taylor). The data set has been cross-checked, proof-read and copy-edited, but is certainly neither exhaustive nor error-free. I will be happy to share the data set with anyone in exchange for any corrections, additions or updates one might have to provide.[2]

We included all publicly available polls that were in the field beginning September 1st, a decision nearly identical to Martin, Traugott and Kennedy's inclusion of polls beginning on Labor Day (2003). This represents a substantially longer time frame than the O'Neill, Mitofsky and Taylor

[2] Only publicly available polls that include sponsor or data collection firm, dates in the field and either sample size or sampling error have been included in the sample. In Colorado, tracking polls with rolling (i.e. overlapping) samples were reported daily; I included only polls without overlap, with one or two exceptions in which a one-day overlap was better than omitting several other days in the field.

(2003) who only included surveys in the field October 20th or later, but substantially shorter than Franklin (2003) who included all surveys beginning in 1999 as long as they identified specifically the names of the eventual major-party candidates. Like Martin et al., I feel that this intermediate time frame is optimal as long as we are sure to take into account the likely biasing effect of real shifts in voter intent over time.

General sample parameters are shown on the bottom row of Table 1. On average, Republican Senate candidates received 52% of the two-party vote in the 232 polls of the sample. The first obvious comparison to be made is to the actual two-party vote, which is not entirely fair because in some of these races there was some real movement in one direction or the other over time. Nonetheless, the two-party vote distribution in the polls is precisely the same as the actual vote to within a tenth of a percentage point.

On average, the polls reported margins of error of 4.1%. This means that 19 times out of 20, or 95% of the time, the poll figures should be within that number of the actual figures, all else being equal. However, this figure is based on the entire sample size, including undecideds and minor-party voters. Since my comparative analysis includes only major-party "decideds" I recalculated margins of error to take that into account resulting in an overall increase in the effective sampling error to 4.4%.[3]

Of course, "all else being equal" is the tricky part. As I will show, below, in several states clear movement took place in the direction of one or the other party during this time frame, and when that occurs it is not fair to judge a September poll solely on the basis of its correspondence to November election results.

The polls in the data set averaged an absolute value 3.0% difference from the actual election results in their state which seems well within this margin of error, but one must recall that 4.4% is not an *average* figure, but rather a figure that represents a *tail* of a distribution beyond which no more than 5% of the sample should fall (2.5% in each direction). This figure of 3.0% error is somewhat higher than that reported by O'Neill, Mitofsky and Taylor (2003), but the difference can easily be explained both by the fact of my longer time frame and the fact that my analysis only includes Senate races while O'Neill et al. also included gubernatorial polls.

While Franklin (2003) and Martin et al. (2003) develop sophisticated tools for testing the accuracy of election polls, I present here the tests actually contained within the probability theory upon which sampling error is based – area under a normal curve, tested with the z-statistic. The most obvious test here is whether the single most basic claim of all surveys holds up: whether only 5% of all cases fall outside of the reported margin of error.

O'Neill et al. report simply that "84% of the polls [in their analysis] differed from the election outcomes by less than their theoretical margin of error." (2003, p. 1) The flip side of this observation is of course that 16% of polls fell outside of the reported margin of error. Thus, while O'Neill et al. report the 84% figure as if it represents an accomplishment by election polls, the reality is different. In my analysis, with its longer time-frame, I find that 25% of surveys fall outside of the reported margin of error (adjusted to take into account the reduced effective sample size). This is even worse than O'Neill et al.'s finding, but even their lower number is

[3] Thanks to Warren Mitofsky (invited address at AAPOR, 2003) and Martin, Traugott & Kennedy (2003) for this critically important observation.

quite damning for election polls – after all, the time frame they use is short enough to rule out real opinion shifts for all but a relatively small percentage of the included polls.

Figure 1: Distribution of Error in Election Polls

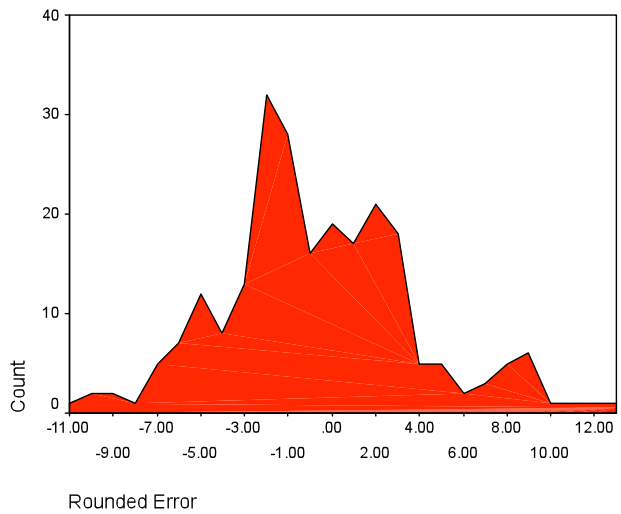


Figure 1, above, shows a distribution of error (difference between Republican percentage of two-party intent in polls and Republican percentage of the two-party vote in the actual vote, rounded to the nearest percentage point). That figure shows quite graphically just how far removed the actual distribution is from the expected normal distribution. The large groupings of polls more than 5% in error on each side show graphically the failure of election polls to meet claimed levels of reliability, even after we have adjusted reported margins of error up to account for decreased effective sample sides (but without adjustments from time-series effects).

In order to conduct additional tests of how the polls in our analysis compare to predicted area under a normal curve it is necessary to convert the average margin of error into a point on the z-distribution. Since 1.96 is the point in the z-distribution beyond which only 5% of cases should be found (or 2.5% in each direction of a two-tailed test) we simply convert various numbers from a scale based on the average effective reported margin of error (4.354%) in units of 1.96. The conversion factor is thus 4.354/1.96 or 2.221.

Using this conversion factor the next test is within what point in the z-distribution 50% of the cases fall. Based on the reported margins of error this should be at 0.67 in the z-distribution, or 1.5 percentage points from the overall mean, with the conversion factor. This compares to my observed average error of 3.0% and O'Neil et al.'s average of 2.4%. But because we are looking at percentages of cases within a particular area beneath a curve we need to look at *median*, not *mean*. The median, at 2.1%, looks a great deal better than the mean, but it is still significantly greater than the 1.5% level where it should lie. In fact, a median of 2.1% is consistent with a point in the z-distribution not of 0.67 where it should be, but rather of 0.95. That figure is larger than the expected 0.67 by a factor of 1.41. When we multiply this factor by the claimed average margin of error of 4.4% we get an observed margin of error of 6.145% (4.4% * 1.41) -- nearly two points higher than the reported margin of error.

The presence of so many cases in the tails of the distribution suggests an additional test using the same methodology. As I observed earlier, 25% of cases fall outside of their reported effective error margins, meaning that 75% lie within that range. The z-score within which 75% of cases

should fall is 1.15, far from the 1.96 figure for 5% of cases. Since 1.96 is larger than 1.15 by a factor of 1.70, this suggests a true margin of error of 7.42% (4.4% * 1.70) – even higher than the test suggested by the distribution's median. If this figure is accurate, then only 5% of all cases should still fall outside of this much higher figure. But even at this much higher level, 18 cases, or nearly 8%, still fall outside of the range. So it would appear that not only does the reported average of 4.2% margin of error substantially understate the true error, *even a much higher margin of 7.4% may still underestimate error in this sample of polls*. While this finding must still be described as very preliminary, due to possible bias from a number of sources not controlled for, it is still fairly breathtaking. The fact that so much is going on in the tails warrants some very careful analysis of outliers, which I undertake below.

Table 1. Comparisons of Republican Senate Candidate Poll Numbers vs. Actual Votes, by State.

State	# of Polls	Average Rep. Poll %	Actual Rep. Vote %	Average Net Difference	Average Error Margin*	Average Abs. Value Difference	Median Abs. Value Difference	% Outside Error Margin	% Predicted Correctly**
Alabama	8	64.8%	59.5%	5.3%	4.3%	6.8%	7.4%	75%	100%
Arkansas	9	46.8%	46.1%	0.7%	4.5%	2.3%	1.9%	11%	89%
Colorado	22	50.4%	52.7%	-2.3%	4.9%	2.4%	1.6%	18%	61%
Georgia	9	46.7%	53.5%	-6.8%	3.9%	6.8%	6.2%	89%	6%
Iowa	12	43.3%	44.8%	-1.5%	4.1%	2.8%	2.1%	25%	100%
Maine	8	63.5%	58.5%	5.0%	4.7%	6.0%	6.4%	63%	100%
Missouri	16	50.6%	50.6%	-0.0%	4.2%	2.4%	2.2%	13%	56%
N. Hampshire	21	51.5%	52.3%	-0.8%	4.3%	3.2%	2.8%	29%	60%
New Jersey	27	45.2%	45.0%	0.2%	4.4%	1.7%	1.5%	4%	98%
No. Carolina	19	55.9%	54.4%	1.5%	4.3%	3.1%	1.5%	16%	100%
Oregon	9	60.1%	58.7%	1.4%	4.3%	2.3%	1.7%	11%	100%
So. Carolina	9	54.5%	55.2%	-0.7%	4.0%	3.2%	4.2%	56%	100%
So. Dakota	19	50.2%	49.9%	0.3%	4.4%	1.7%	1.7%	0%	47%
Tennessee	19	56.6%	54.2%	2.4%	4.4%	2.9%	1.7%	26%	100%
Texas	25	53.8%	56.0%	-2.2%	4.1%	3.4%	2.3%	32%	80%
Total	232	52.1%	52.2%	-0.1%	4.4%	3.0%	2.1%	25.0%	79%

*Average of reported margin of error adjusted to account for reduced sample size due to omission of undecideds and minor-party voters.

**Polls with a "tie vote" count as half a correct prediction.

Breaking It Down by State

The other rows of Table 1 show a great deal of variance among the states with regard to poll accuracy. For example, polls in Alabama and Maine, with safe Republican incumbents nonetheless overestimated Republican vote by 5% (and erred by medians of 7% and 6% respectively); In Colorado it was the Democratic candidate whose fortunes were overestimated, by 2.3%, making the difference between what appeared to be a very tight race in the polls and a relatively easy Republican victory. Similarly, In Tennessee the Republican vote was overestimated by around 2.4% while in Texas the Democratic vote was overestimated by a similar margin. In Georgia, the vote for Democratic incumbent Cleland seems at first glance to have been systematically over-estimated, but in that case the polls actually show his numbers dropping steadily as election day approached. (See the scatter-plots in Figure 2.)

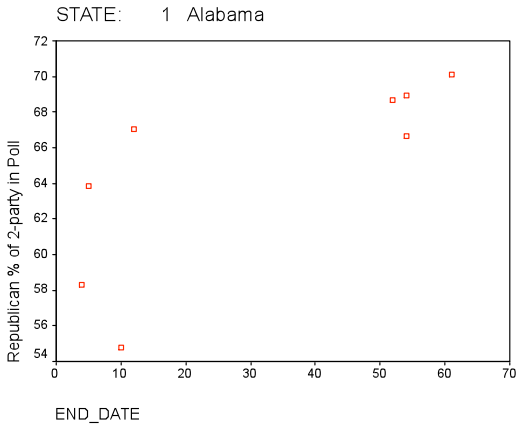
Figures 2a-f are simple scatter-plots showing Republican percentage of the two party vote in polls on the y-axis against the end-date of the poll on the x-axis.[*] The states shown in these scatterplots are those showing clear movement over time. For example, in the Arkansas plot (Pearson's R = -.68), Tim Hutchinson led in only two polls in September; every poll from that point on had Democratic challenger Pryor ahead, and by increasingly large margins quite consistent with his eventual decisive victory. The Georgia plot (R = .59), while not quite so clear-cut, shows Democratic incumbent Max Cleland largely ahead early on, but challenger Saxby Chambliss pulling even and slightly ahead by the end, with momentum on his side. Similarly, Missouri (R = .58) shows a Republican trend (where the race eventually went), while New Hampshire (R = -.46) and North Carolina (R = -.41) show declining Republican strength leading to a tightening of the race, although Republicans eventually won both states as well. Interestingly, Alabama shows a pro-Republican trend over time (R = .75) that was evidently imaginary – Alabama polls became less accurate, predicting Republican landslides by ever greater margins, as the election drew closer.

For comparison's sake, Figures 3a-e show similar scatter-plots for states that showed no such systematic movement over time: Colorado (R = -.01), Oregon (R = -.15), South Dakota (R = .08), Tennessee (R = -.18) and Texas (R = -.08). Tennessee and Texas are of particular interest in this regard, since in both states some analysts perceived a movement toward the Democratic candidate (see Neumann, 2002), while in the end, the Republican won by large margins (although in Tennessee the polls overestimated Republican strength, while in Texas they underestimated it).

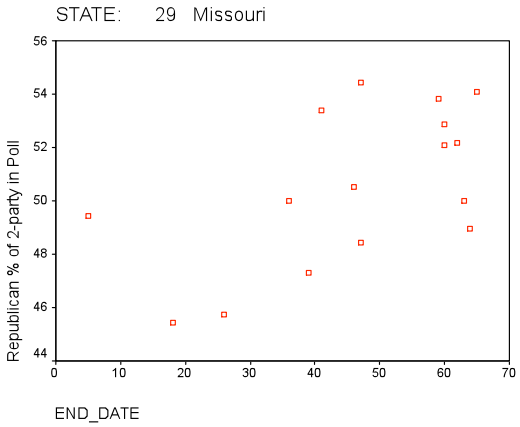
In these states, along with others which showed no systematic movement over time – especially Maine and South Carolina – it would seem much more fair to look at the extent to which the poll results differed from election results by more than the polls' reported margins of error (as shown in Table 1). As before, this is a mixed bag: in South Dakota not one poll fell outside of the margin of error around the election results. On the other hand, 56% in South Carolina (5 of 9 polls) and a whopping 63% in Maine (five of eight polls) were outside the margin of error. Overall, with the exception of Georgia, there does not seem to be much correspondence between the extent to which polls fell outside of the margin of error and the extent to which systematic movement occurred over the time period of the analysis.

[*] Dates are numbered so that September 1st is "1;" September 30th is "30," October 1st is "31," etc. through November 4th, which is "65."

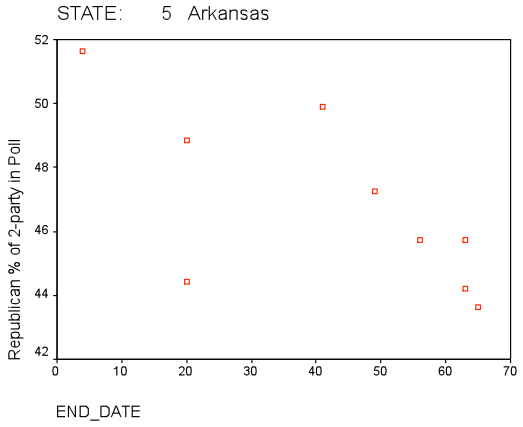
Figures 2a-f. Republican Percentage in Senate Polls in States With Movement Over Time, September 1, 2002 – November 4, 2002



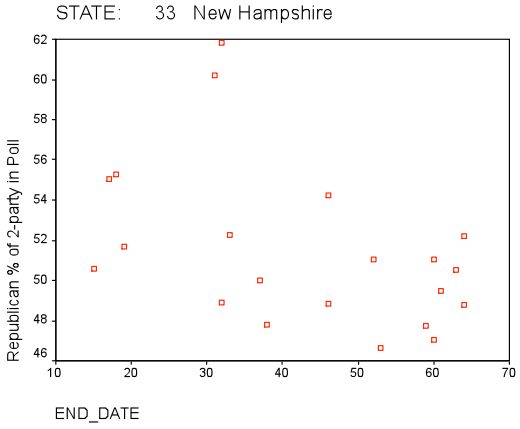
Alabama R = .754 (Sig. = .031)



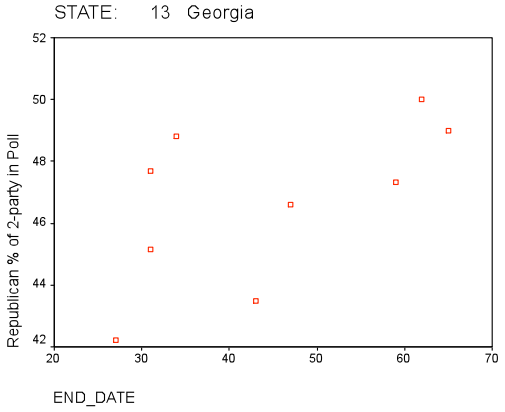
Missouri R = .580 (Sig. = .018)



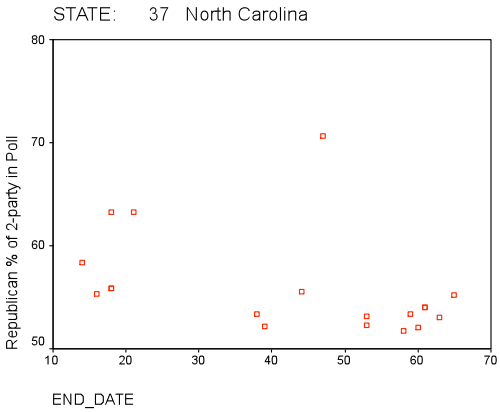
Arkansas R = -.681 (Sig. = .043)



New Hampshire R = -.461 (Sig. = .035)

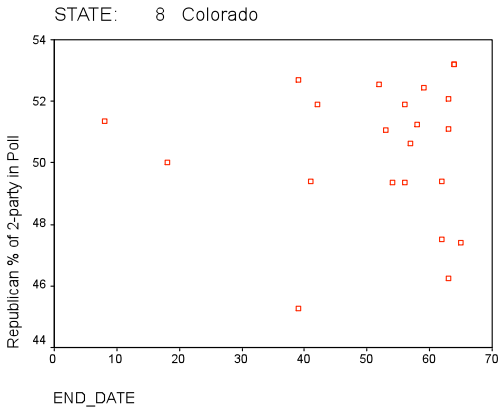


Georgia R = .592 (Sig. = .093)

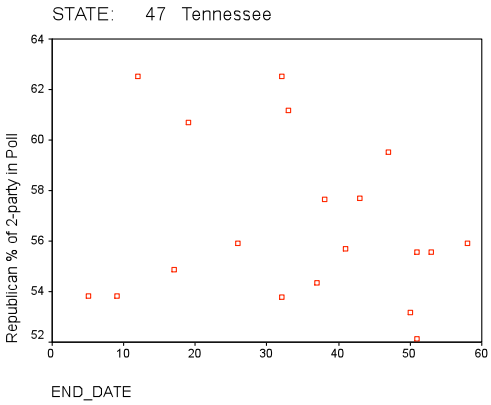


North Carolina R = -.412 (Sig. = .080)

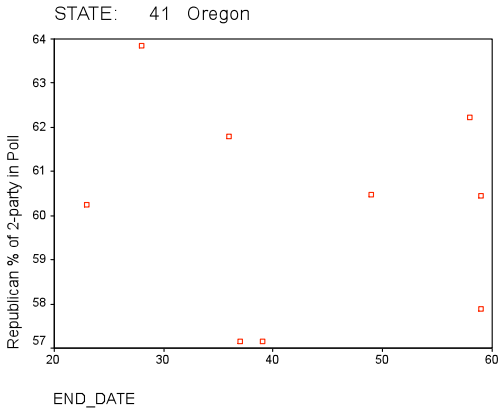
Figures 3a-e. Republican Percentage in Senate Polls in States With No Systematic Movement Over Time, September 1, 2002 – November 4, 2002



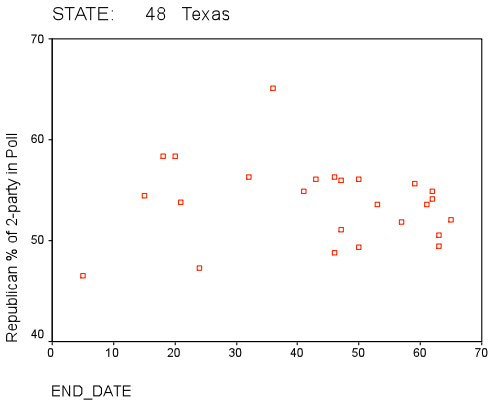
Colorado R = -.006 (Sig. = .977)



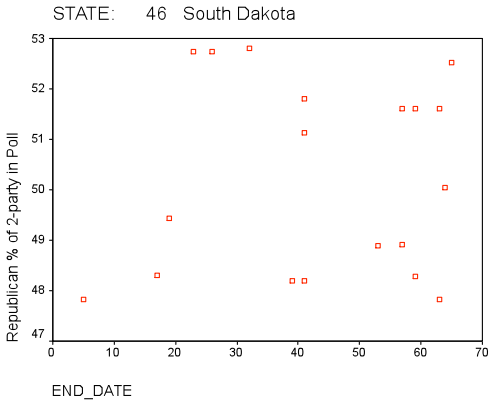
Tennessee R = -.178 (Sig. = .467)



Oregon R = -.145 (Sig. = .709)



Texas R = -.079 (Sig. = .707)



South Dakota R = .080 (Sig. = .746)

Looking at level of error (again as measure simply by looking at the difference between Republican strength in 2-party vote intent in the poll and the actual vote) produces results sometimes similar and sometimes different from a simple look at Republican strength by itself. Figures 4a-f and 5a-e show scatterplots of error over time while Table 2 show the correlation between end-date of the survey (the last day the survey was in the field) and level of error.

All else being equal we would expect that error should increase with greater distance in time between the poll and the election and overall the data bear out this expectation, albeit not as consistently and not by as large a measure as we might expect. Looking at all 232 polls, the overall correlation is -.15, a small-to-moderate size correlation that is statistically significant (P =.05). Eleven states have negative correlations to four with positive correlations. Only one correlation, Alabama's large positive one, achieves statistical significance, but this is due largely to the very small sample sizes in individual states.

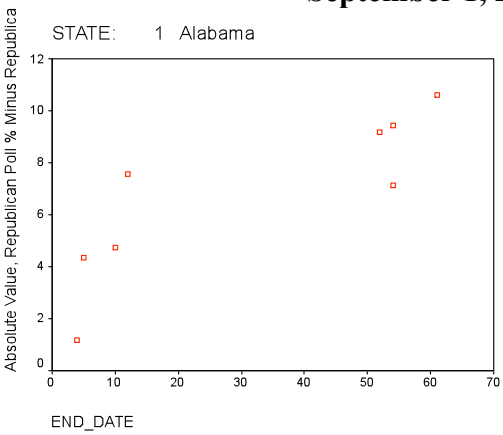
Five of the six states presented in Figures 2a-f – the ones in which polls accurately measured movement over time – have negative correlations between error and date of the survey, with Arkansas (R = -.66) and Georgia (R = -.59) topping the list. Missouri and North Carolina have smaller correlations (-.262 and -.225 respectively), while New Hampshire's is negligible. This is because in New Hampshire most polls showed the Democratic candidate, Jeanne Shaheen, gaining strength right up until a week or so before the election, but then Republican John Sununu benefiting from a late surge. That makes the graphs of both Republican strength and error non-linear – resembling an "L" in the first case and a "J" in the latter.

Not surprisingly, the states identified as showing little or no net movement over time show smaller time effects. Of the five states identified as such in Figures 3a-3, Oregon and South Dakota show small-to-moderately large correlations (-.29 and -.24 respectively); Tennessee and Texas show small correlations (-.13 and -.18 respectively) and Colorado's figure is negligible.

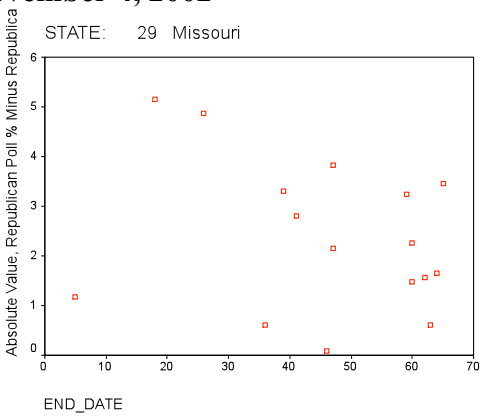
Table 2. Correlation between Poll Error and Date of Interview, by State (In order of Direction and Size of Correlation)

State	Number of Polls	Correlation	Significance
Arkansas	9	-.662	.052
Georgia	9	-.592	.093
New Jersey	27	-.338	.084
Maine	8	-.305	.462
Oregon	9	-.294	.442
Missouri	16	-.262	.327
South Dakota	19	-.243	.316
North Carolina	19	-.225	.293
Texas	25	-.177	.398
Tennessee	19	-.131	.594
New Hampshire	21	-.057	.805
Colorado	22	.040	.858
Iowa	12	.142	.659
South Carolina	9	.404	.281
Alabama	8	.843	.009
Total	232	-.154	.019

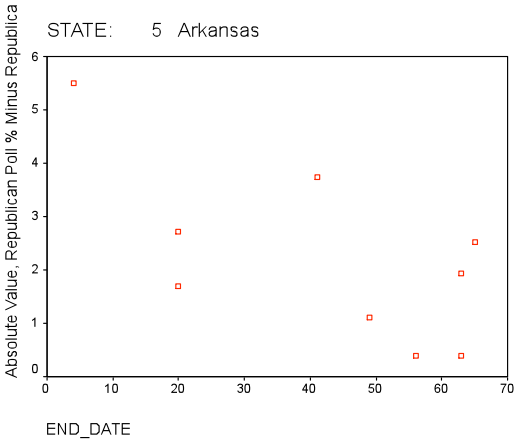
Figures 4a-f. Absolute Value Error in Senate Polls in States With Movement Over Time, September 1, 2002 – November 4, 2002



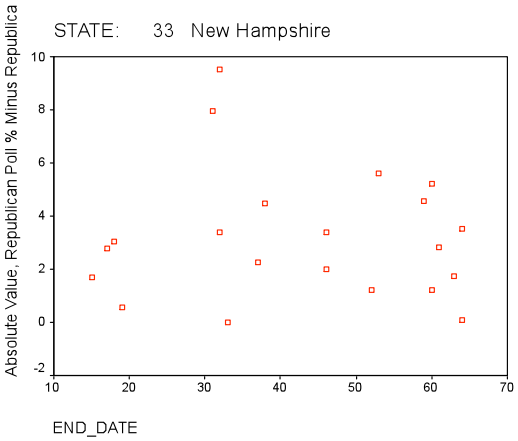
Alabama R = .843 (Sig. = .009)



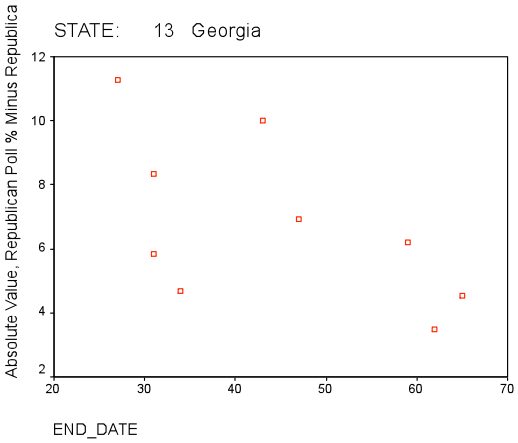
Missouri R = -.262 (Sig. = .327)



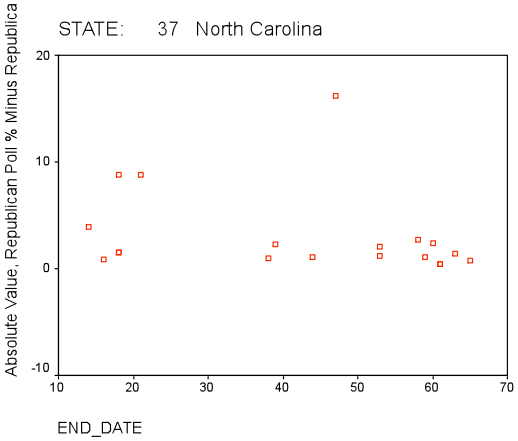
Arkansas R = -.662 (Sig. = .052)



New Hampshire R = -.057 (Sig. = .805)

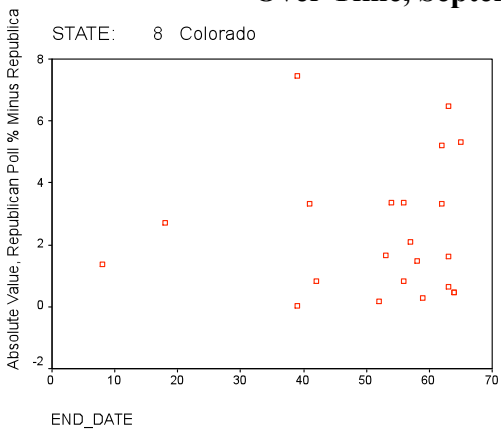


Georgia R = -.592 (Sig. = .093)

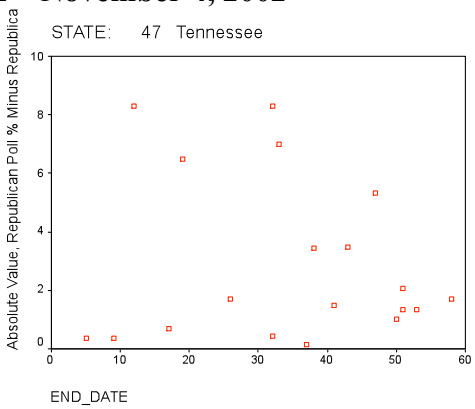


North Carolina R = -.255 (Sig. = .293)

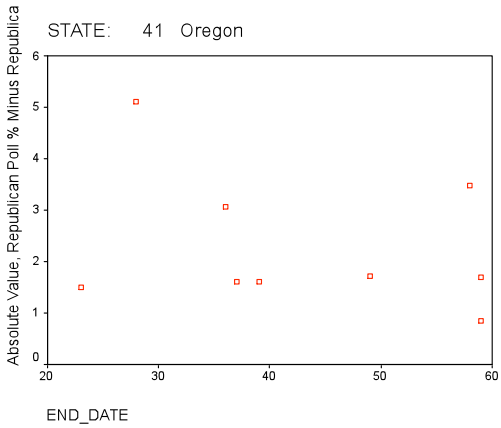
Figures 5a-e. Absolute Value Error in Senate Polls in States With No Systematic Movement Over Time, September 1, 2002 – November 4, 2002



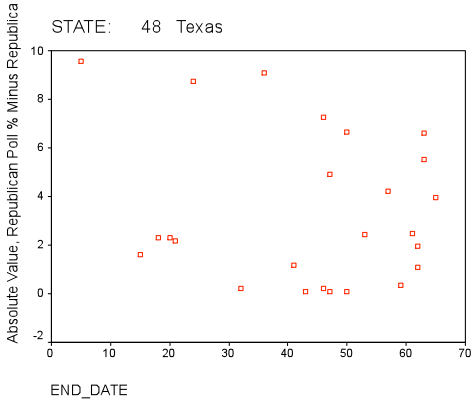
Colorado R = -.040 (Sig. = .858)



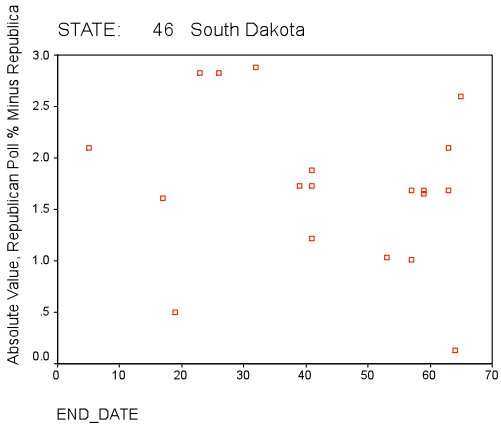
Tennessee R = -.131 (Sig. = .594)



Oregon R = -.294 (Sig. = .442)



Texas R = -.177 (Sig. = .398)



South Dakota R = -.243 (Sig. = .316)

Thus, for this set of races in 2002 using the date of the survey as a predictor of its accuracy is not as clear-cut as it might normally be. Overall the correlation is negative, as expected, but not strongly so, and important exceptions exist, weakening the overall effect and making it more difficult to control for date-of-survey effects. A closer look of this relationship will be the subject of future research.

Predictive Accuracy

So far the polls in my data set have performed well by some measures (average vote estimates), but poorly on others (median absolute value of the difference between the poll and the outcome) and very poorly on one more (percentage of cases outside of the margin of error). But what about the "money" question – how well did these polls do at predicting winners?

Here we have no statistical method for measuring success (and sampling methodology makes no claims here) so we must fall back on entirely subjective judgments. Overall, these polls got the winner right 79% of the time, varying from a low of 6% (one tie out of 9 polls) in Georgia to highs of 100% in several states. But how important a measure of poll accuracy is this? After all, polls in Alabama were all over the place, but all correctly predicted a Republican victory; on the other hand, only 47% of polls got it right in South Dakota, but all were within the margin of error and were correct in predicting a close race decided by only a few hundred votes. If we leave out the states that were never in doubt (Alabama, Iowa, Maine, North Carolina, Oregon, South Carolina and Tennessee) polls in the remaining states only got it right 67% of the time, but is that really so bad? Colorado, Georgia, Missouri, New Hampshire and South Dakota were close races and were in play until the end. Georgia, in particular, probably really did lean toward Cleland until close to the end of the campaign and the late polls picked up that movement.

The Accuracy of Partisan Pollsters

An additional question of interest is the matter of whether partisan polls are as accurate as the more commonly published nonpartisan media or academic polls. Happily, this is a far easier question to address than the more general question of poll accuracy. The first question is whether partisan polls have a systematic bias in favor of their party's candidate. A look at Table 2 shows that this does indeed seem to be the case: in polls conducted by Democratic pollsters, Democrats did 3% better than the eventual election outcome, while Republican candidates were similarly blessed with an additional 2% by Republican pollsters (with just a slight net pro-Democratic bias among nonpartisan polls). As we will see below, this bias remains in a multivariate regression model. Partisan polls also had a larger absolute value average difference of 4% compared to 3%, larger median absolute value differences of 3% compared to 2%, and had far poorer prediction rates than nonpartisan polls.

It seems unlikely here that this is a result of partisan pollsters providing their clients with bad data. If that were the case, their clients would not spend so much money on polls. Rather, their polling data remains proprietary in all cases except those in which the campaign feels it would gain strategically from publishing the results. Since it is always better to be perceived as winning by more or losing by less, it is not surprising that the partisan polls that see the light of day tend to paint a rosier picture than the reality for the candidate who sponsored the poll.[4] If this is the

[4] Thanks to Democratic pollster Mark Mellman for this observation (personal conversation at the Annual Meeting of the American Association for Public Opinion Research, May, 2003).

case, then both journalists and academics should be particularly careful when using the selectively published results of partisan polls. (More on this below.)

Lightning Rods: Zogby and Survey USA

While election polling in general took a lot of heat after the 2002 elections, two polling operations came under particularly heavy fire – Zogby International for larger than usual numbers of missed calls, and Survey USA for its controversial use of automated telephone interviewing systems. Since we have 30 or more polls from each of these firms in our data set, it would seem a worthwhile exercise to test whether the criticism to which they have been subjected is fair or not.[5]

John Zogby hit the jackpot in 1996 when his firm was the only one not to overestimate Clinton's margin of victory. Zogby's accurate prediction that year quickly gave him credibility (especially among Republicans, although he is nonpartisan) and made him a star. His numbers in 2000, however, were not so close, and he had a couple of real embarrassments, such as his poll that showed Hillary Clinton behind Rick Lazio just before the election. Within the polling community, it was known that Zogby weights his samples not only by demographic variables, but also by partisanship, predicting what percentage of the electorate will be of each party, a technique that some think comes close to cooking the books, which is why it is discouraged by the National Council on Public Polls (O'Neill, 2002a; see Crespi, 1988, for a discussion of the prevalence of this technique in the 1980s). (It turned out that the reason he was so far off in the Clinton-Lazio race was that he was using percentages of Democrats in the electorate based on the last Senate race in New York State – the 1998 off-year election – rather than the last presidential election year!)

[5] The analysis conducted for the Nation Council on Public Polls (O'Neill et al, 2002) omitted Survey USA surveys without comment. The apparent reason, judging from one co-author's public comments at the 2003 AAPOR meeting (Warren Mitofsky) was a discomfort with that firm's use of automated interviewing techniques rather than human interviewers. In my opinion this omission is not justified in the way that omissions of surveys using non-probability sampling techniques would clearly be. SurveyUSA uses probability sampling techniques that fall well within the norms of legitimate survey research. Individuals are free to doubt, as I did, that automated interviewing could not be as effective or reliable as interviewing by human interviewers, but far from suggesting omission of SurveyUSA from the data sets, these sorts of concerns argue in support of including them so that we can actually test how they compare to other operations using more traditional methodologies. In that light, and from a social scientific perspective in general, O'Neill et al's off-hand decision to omit all Survey USA polls is really quite shocking. As the reader can see, that methodology does indeed stack up quite well compared to traditional interviewing.

Table 3. Comparisons of Republican Senate Candidate Poll Numbers vs. Actual Votes, by Partisanship of Poll, Zogby & Survey USA, and by Week.

	# of Polls	Average Rep. Poll %	Actual Rep. Vote %	Average Net Difference	Average Error Margin	Average Abs. Value Difference	Median Abs. Value Difference	% Outside Error Margin	% Predicted Correctly
Democratic	26	51.0%	54.2%	-3.2%	4.4%	3.9%	3.0%	31%	65%
Nonpartisan	180	52.0%	52.0%	-0.0%	4.3%	2.8%	1.9%	24%	84%
Republican	26	53.2%	50.9%	2.3%	4.6%	3.7%	2.9%	27%	60%
Zogby	35	51.3%	51.7%	-0.4%	4.6%	3.1%	2.6%	20%	76%
Survey USA	30	50.1%	51.8%	-1.7%	4.1%	2.2%	1.8%	20%	85%
Others	167	52.6%	52.7%	-0.1%	4.4%	3.1%	1.9%	27%	78%
Weeks 1*	68	49.8%	51.1%	-1.3%	4.2%	2.5%	1.7%	21%	68%
Week 2	31	53.5%	52.7%	0.8%	4.4%	2.8%	1.9%	19%	87%
Week 3	28	52.8%	52.2%	0.6%	4.3%	3.3%	2.0%	29%	82%
Week 4	27	51.2%	52.1%	-0.9%	4.4%	2.5%	1.9%	11%	74%
Week 5	26	52.4%	51.1%	1.3%	4.3%	3.6%	2.9%	35%	75%
Week 6	7	49.5%	52.5%	-3.0%	4.3%	5.8%	5.1%	71%	43%
Week 7	25	54.9%	53.2%	1.7%	4.6%	3.2%	2.7%	24%	90%
Week 8	11	55.7%	54.0%	1.7%	4.4%	4.2%	3.9%	36%	100%
Week 9	9	53.0%	53.7%	-0.7%	4.4%	2.9%	1.4%	33%	67%
Total	232	52.1%	52.2%	-0.1%	4.4%	3.0%	2.1%	25.0%	79%

*Weeks *before* the election.

Zogby's last published polls in 2002 included some that not only called the winner incorrectly, but did it by a fairly large amount. His last poll in Colorado had Strickland up by 5 points; he had Cleland ahead in Georgia, Thune up by 5 in South Dakota, and was one of several pollsters who saw the Texas race as close. His last poll in Arkansas had Pryor winning by 13 points and he had Jean Carnahan behind by 8 points in Missouri. So even when he had the winner right he was often off the mark in terms of the margin. These were balanced by more accurate final polls in New Jersey, North Carolina and Tennessee.

Be that as it may, it is easy to test whether Zogby performed worse than other polling operations in 2002, or whether he just had bad luck in having his worst outings in the most highly publicized final polls. Looking at Table 3, above, it is first of all clear that Zogby's polls showed no net partisan bias. In terms of the average absolute value difference between his polls and the election results, he is right on the sample average (which of course is not saying much). Interestingly, his median absolute value error, at 2.6%, is substantially higher than the overall figure. His percentage outside the margin of error, at 20%, while still far higher than it should be, is well below both the overall average and the average for nonpartisan polls. Finally, despite his high-profile errors in the final polls, Zogby's overall record of 76% correct predictions was only slightly lower than the 79% average, albeit well below the 84% average for nonpartisan polls.

Looking at Survey USA, they too stuck with Carnahan in Missouri, and had both Carolinas much tighter than they actually were. But they were right in Colorado, and had the margins closer elsewhere. Interestingly, they were not in as many close states as Zogby, so the fact that 85% of their polls had the correct winner should not be taken too far. Unlike Zogby, or nonpartisan pollsters in general, Survey USA did show a net pro-Democratic bias of around 2 points, but by every other measure they performed as well as or better than other nonpartisan firms. Thus, as much as academic survey researchers may have wished to see Survey USA under-perform the field, they clearly did not, and may have actually done better than average.

Multivariate Analysis

To test whether the differences among categories of pollsters hold up when other important factors are held constant, I conducted a simple OLS regression analysis with Republican percentage of the two-party support in the poll as the dependent variable; the results appear in Table 4, below.[6] Dummy variables for all states except New Hampshire are included in order to take into account characteristics of the races in each state (New Hampshire simply had the result closest to the average results). Similarly, dummy variables for each week in the field except the last are included in order to take into account even temporal effects that might not be entirely consistent across the time frame of the analysis.

Looking first at the coefficient for the type of poll, the bias of partisan polls remain, at 4% for Democratic pollsters and 2% for Republican pollsters, with both effects strongly significant statistically. So the caveat about reading with caution the results of published polls from partisan sources remains as strong as before. The finding that Survey USA had a net pro-Democratic bias of around 2% also has held up in the regression coefficients, as did the finding that Zogby polls showed no such bias.

Coefficients for the week variables mostly reflect the comparison with the omitted week (1 week before the election), which for some reason, showed a strong, and apparently inaccurately measured movement toward the Democrats.[7] All week variables have positive coefficients, with 3, 5 and 7 weeks before the election showing especially large (and statistically significant) differences with the omitted final week. Similarly, the coefficients for the states mostly just reflect the comparison with the omitted New Hampshire.

Overall statistics for the regression are quite strong, with an adjusted R^2 of .745, although given the presence of the state dummy variables that is hardly surprising or noteworthy.

[6] Because the dependent variable here is Republican percentage in the polls, with no reference to the actual vote the model does not judge the accuracy of poll results, only net partisan direction.

[7] Interestingly, while Zogby polls showed no net bias towards the Democrats, his polls during the last week did, and his polls constituted a large percentage – 17 of 68 – of all polls in the data set in that period.

Table 4. Multivariate Regression Analysis of Republican Poll Percentage

Variable	Unstandardized Coefficients	Std. Error	Standardized Coefficients	t-ratio	Sig.
(Constant)	50.525*	.783		64.548	.000
Democratic Pollster	-4.174*	.758	-.210	-5.507	.000
Republican Pollster	1.925*	.770	.095	2.501	.013
Zogby	-.228	.710	-.013	-.322	.748
Survey USA	-2.237*	.716	-.120	-3.126	.002
9 weeks before election	.725	1.258	.022	.576	.565
8 weeks before election	1.284	1.098	.043	1.169	.244
7 weeks before election	1.735*	.793	.086	2.187	.030
6 weeks before election	.022	1.327	.001	.016	.987
5 weeks before election	3.101*	.798	.156	3.888	.000
4 weeks before election	.782	.747	.040	1.047	.296
3 weeks before election	1.934*	.777	.100	2.490	.014
2 weeks before election	.141	.755	.008	.187	.852
Alabama	13.044*	1.350	.401	9.665	.000
Arkansas	-5.038*	1.405	-.146	-3.587	.000
Colorado	-.041	1.018	-.002	-.041	.968
Georgia	-4.621*	1.302	-.142	-3.550	.000
Iowa	-7.842*	1.186	-.277	-6.613	.000
Maine	12.189*	1.364	.354	8.933	.000
Missouri	-.147	1.108	-.006	-.132	.895
New Jersey	-6.010*	.951	-.307	-6.316	.000
North Carolina	5.626*	1.049	.246	5.364	.000
Oregon	8.192*	1.286	.252	6.370	.000
South Carolina	4.744*	1.353	.146	3.506	.001
South Dakota	-1.071*	1.053	-.047	-1.018	.310
Tennessee	5.561*	1.057	.243	5.264	.000
Texas	3.412*	.980	.168	3.483	.001

Dependent Variable: Republican % of 2-party support in Poll
R^2 = .773, Adjusted R^2 = .744, SEE = 3.183
* Statistically significant at at least .05 level.

Table 5, below, shows regression coefficients for an analysis with survey error (again, difference between the poll and the election results) as the dependent variable, with error once again simplistically measured as the absolute value difference between Republican poll numbers and the actual vote. All variables from the previous model are included along with one additional variable, survey sample size.

First, looking at the overall model statisics, the R^2 is .308 (adjusted R^2 = .216), indicating that most of the variance in survey error is caused by factors not included in the model. However, if sampling error is working precisely as it should, the null hypothesis would be that error would be entirely random and the expectation would be an R^2 of 0 with no statistically significant coefficients (except of margin of error). Clearly, that is not the case.

Table 5: Multivariate Regression Analysis of Election Poll Error

Variable	Unstandardized Coefficients	Std. Error	Standardized Coefficients	t-ratio	Sig.
(Constant)	5.772*	1.828		3.158	.002
Democratic Pollster	.346	.581	.040	.596	.552
Republican Pollster	1.269*	.601	.143	2.113	.036
Zogby	.998	.558	.130	1.789	.075
Survey USA	-.709	.555	-.087	-1.278	.203
9 weeks before election	-.087	.969	-.006	-.090	.929
8 weeks before election	.930	.854	.072	1.088	.278
7 weeks before election	.278	.624	.031	.446	.656
6 weeks before election	3.561*	1.027	.222	3.466	.001
5 weeks before election	1.253*	.618	.144	2.028	.044
4 weeks before election	.235	.580	.027	.406	.685
3 weeks before election	.797	.599	.094	1.329	.185
2 weeks before election	.029	.583	.004	.050	.960
Alabama	3.254*	1.034	.229	3.146	.002
Arkansas	-1.056	1.076	-.070	-.981	.328
Colorado	.043	.818	.005	.052	.958
Georgia	2.529*	1.016	.178	2.489	.014
Iowa	-.542	.911	-.044	-.595	.552
Maine	3.221*	1.052	.214	3.061	.003
Missouri	-.945	.851	-.087	-1.111	.268
New Jersey	-1.276	.733	-.149	-1.741	.083
North Carolina	.074	.804	.007	.092	.927
Oregon	-1.203	.986	-.084	-1.220	.224
South Carolina	.389	1.040	.027	.375	.708
South Dakota	-1.685*	.807	-.168	-2.089	.038
Tennessee	-.431	.810	-.043	-.532	.595
Texas	.081	.755	.009	.107	.915
Margin of Error	-.759	.419	-.133	-1.811	.072

Dependent Variable: Absolute Value, Republican Poll % Minus Republican Vote %, 2 party
R^2 = .308, Adjusted R^2 = .216, SEE = 2.439
* Statistically significant at at least .05 level.

Looking at some of the coefficients in the model, first of all, it is interesting that despite the greater net partisan bias found among Democratic polls, the level of error is higher among Republican polls, with more than an extra point added to error, a statistically significant effect. The error level was around a point higher in Zogby polls and a point lower in Survey USA polls than among other nonpartisan pollsters, but neither effect is statistically significant.

The week variables don't show much; only two have statistically significant coefficients, and the larger of those (and by far the largest for any week) is for 6 weeks before the election, which had only 7 cases. Nonetheless, the overall importance of including the week variables is demonstrated when one conducts the same regression analysis omitting them. The regression sum of the squared errors (SSE) decreases from 539.74 to 447.0, a difference of over 90, far higher than the critical X^2 value of 15.5 (.05, 8 DF; analysis not shown in Table 5)[8]. Again, these

[8] This test is explained in Hanushek & Jackson (1977), pp. 125-126.

results are consistent with earlier findings presented here regarding the complicated nature of attempting to control for time effect.

As in the previous analysis, state values to a large extent reflect differences between the particular state and the omitted New Hampshire, which was somewhat above average in its levels of error. Of particular interest here are the three states with substantially (and significantly) higher error – Alabama, Georgia, and Maine – and the one state with substantially (and significantly) lower error, South Dakota. Analysis of what could be behind these differences among states will await further analysis.

In analysis not shown here, I attempted to test two additional hypotheses: (1) error will be higher in states with fewer polls; and (2) error will be higher in polls conducted by pollsters who do not conduct a great many polls. The first hypothesis based on the observation that when few polls exist in a state, it is more difficult for pollsters to spot errors or trends and take them into account. In particular, it is hard to miss the fact that the three states with by far the highest average error (Alabama, Georgia and Maine) only had eight or nine polls each. Looking at bivariate correlations, the number of polls in the state had a Pearson's R of -.260 and a dummy variable in which states with fewer than 10 polls were coded "1" had an R of .284 (both correlations are statistically significant at beyond the .01 level.) Unfortunately, both versions of this variable brought in levels of multicollinearity with the string of state dummy variables that were too high to include all together in the same model. When included in regressions without the state dummy variables, these variables had large, statistically significant, coefficients in the expected direction. In effect, the information about the small number of polls was already wrapped into the state dummy variables anyway.

Moving on to the second hypothesis, that polls conducted by polling companies that don't conduct a lot of election polls should have higher error, this variable ran into some difficulties too. The main problem, as I was able to recognize informally looking at the data set, was that a lot of the firms that only had one or two polls in the sample are actually very experienced firms that do a lot of election polling. This had the effect of both reducing the variable's effect and increasing the error associated with it. As a result, the Pearson's R between this variable and error was only -.121 and not statistically significant. Similarly, when it was included in the regression model it produced a coefficient of around -.05 that was not statistically significant. The main impact of including this variable in the model was to increase the coefficients of the Zogby and SurveyUSA dummy variables, since both firms conducted large numbers of polls included in the sample. Since the coefficient was not significant I decided it was better to leave it out.

Conclusion

This examination of survey data shows mixed results as to the reliability of election polls in 2002. By some measures the polls seem to be quite reliable, while by others – most notably the percentage of polls that fall outside their own published margin of error compared to the election results – they perform quite poorly. The most striking preliminary finding is that some evidence suggests support for the hypothesis that reported margins of error may be vastly understated. While adjusted effective reported margin of error is 4.4%, actual observed error by one measure appears to be 6.1%; by another it could be 7.4%; by yet another it could be higher still. On the strictly subjective measure of predictive accuracy, they are also a mixed bag.

One important, although not surprising, finding is that partisan polls reflect a partisan bias; a consumer of such polls should always keep this in mind and add or subtract a few points as appropriate. Most of us had doubtless assumed this, but "now we know."

The other fairly strong finding here is that despite the criticism to which they were subjected, Zogby International and Survey USA performed at roughly the same level of other nonpartisan polling organizations in 2002, with Zogby performing a bit below average and SurveyUSA a bit above it. Given the level of suspicion in which many survey researchers hold that firm, this find may appear to be a ringing endorsement indeed.

Much more remains to be done here, including not least of all, further copy editing and expanding the data set. I might expand the data set to include gubernatorial polls, or to encompass a longer time frame. I will also be applying more sophisticated analytical techniques including time series analysis and others in an effort to come up with the still-elusive definitive answer to the question as to whether election polls perform as advertised.

The presence of such large number of surveys at the tail of the reliability distribution suggests a closer look at outliers as well. It remains quite possible that overall results could be biased either by a small number of particularly unreliable firms or by large numbers of firms that don't conduct election polls regularly and that are thus subject to higher levels of human error.

While Martin et al. (2003) and Franklin (2003) have a great deal to offer with their sophisticated multivariate methodologies (and will doubtless prove very useful), I believe that the analysis presented here in many ways is more appropriate. After all, pollsters make no claims about what the natural logorithm of the odds ratio should be (as in Martin et al.), but they do make claims about what percentage of cases should fall within particular distances of the election outcome. The question still remains, "reliable compared to what?" but I do believe this paper makes significant progress in answering that question.

REFERENCES

Baldassare, Mark, Mark DiCamillo and Susan Pinkus. "Polling in the Governor's Race in California, 2002." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, May 16-19 2003, in Nashville, TN.

Crespi, Irving. 1988. *Pre-Election Polling: Sources of Accuracy and Error.* New York: Russell Sage.

Franklin, Charles. "Polls, Election Outcomes and Sources of Error." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, May 16-19 2003, in Nashville, TN.

Green, Donald P. and Alan S. Gerber. "Enough Already with Random Digit Dialing: Using Registration-Based Sampling to Improve Pre-Election Polling." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, May 16-19 2003, in Nashville, TN.

Harrison, Chase H. "Coverage Bias in Telephone Samples of Registered Voters." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, May 16-19 2003, in Nashville, TN.

Jackson, John E. and Eric A. Hanushek. *Statistical Methods for Social Scientists.* 1977. San Diego: Harcourt Brace.

Martin, Elizabeth A., Michael W. Traugott, and Courtney Kennedy. "A Review and Proposal for a New Measure of Poll Accuracy." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, May 16-19 2003, in Nashville, TN.

McDonald, Michael. "Voting-Age and Voting-Eligible Population Turnout Rates" Fairfax, Virginia: George Mason University. (http://elections.gmu.edu/VAP_VEP.htm)

Morin, Richard. "Smackdown in Maryland: RBS versus RDD." 2003. *Public Perspective* 14, Number 1; 7-9, 41.

Neumann, Johanna. "Looking to History, Pundits Never Saw This One Coming." Los Angeles Times, November 7, 2002.

O'Neill, Harry, Warren Mitofsky and Humphrey Taylor. "National Council on Public Polls Polling Review Board Analysis of the 2002 Election Polls." National Council on Public Polls (NCPP) press release, December 19, 2002.

O'Neill, Harry, Warren Mitofsky and Humphrey Taylor. 2002a. "The Good and Bad of Weighting Data," a statement by the National Council on Public Polls Polling Review Board.

Traugott, Michael W. and Paul J. Lavrakas. 2000. *The Voter's Guide to Election Polls, 2nd Ed.* New York: Chatham House.