

## **A Probability-Theory Based Test of the Reliability of Election Polls**

**Joel David Bloom**

The University at Albany, SUNY  
jdbloom@albany.edu

**Jennie Elizabeth Pearson**

The University of Nebraska, Lincoln  
jennie.pearson@gmail.com

**Abstract:** We analyze a data set of 344 polls from presidential elections in 45 states. Using a methodology first developed in an earlier paper (Bloom, 2003) we examine the reliability of the polls with the yardsticks provided by probability theory on which the concept of sampling error is based. Unlike the off-year election Senate polls examined in that paper, we find that the presidential polls are every bit as reliable as advertised. While observed error from election results in the 2002 Senate sample substantially exceeded the polls' reported margins of error, we find that the 2004 statewide presidential polls varied from the election results precisely as predicted by probability theory. On the strictly subjective measure of predictive accuracy, these surveys also performed better than the 2002 Senate polls. Another interesting finding, consistent with the previous paper, is that despite criticism over the use of automated interviewers, Survey USA and Rasmussen Reports provided accurate and reliable data, consistent with other polling organizations.

*NOTE: After presenting this paper, the authors realized they had should have accounted for rounding error in their calculations, which would have had the net effect of increasing the expected error for most polls. This was corrected in the subsequent work, "Reliable Compared to What? A Probability-Theory Based Test of the Reliability of Election Polls," in Elections and Exit Polling, Scheuren and Alvey, eds. (2008). Wiley: NY.*

This paper was prepared for presentation at the 2005 Annual Meeting of the American Association of Public Opinion Research, May 2005, in Miami, Florida. Comments and suggestions are welcome. We will be happy to share the data set on request in exchange for any corrections or additions you may have.

## **Introduction: Reliable Compared to What?**

Pollsters were the favorite media punching bag after the November 2002 elections, in which Republicans won somewhat more Senate races and many 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.

The strongest criticism came from columnist Ariana Huffington, who has launched a “crusade” against polls:

I'm still trying to figure out who had a more wretched Election Night 2002, the Democratic Party or America's pollsters. While Democrats lost control of the Senate, they will live to fight another election day. Pollsters, on the other hand, in losing what scraps of credibility they had, may – with a little help from the public – find their entire profession obsolete, gone the way of chimney sweeps, organ pumpers, and those guys who used to make buggy whips. (Huffington, 2002)

Indeed, as Huffington pointed out, there appeared to be some major misses by the pollsters in the Senate and gubernatorial races. 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. For the 2004 analysis another automated survey practitioner is added to the mix. So we will also take up the question of whether polls conducted by these firms in particular, but other large firms as well, 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.

In a previous work (Bloom, 2003) we attempted to answer these questions using a data set of 232 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 complicating runoff and Minnesota, where incumbent Paul Wellstone died less than two weeks before the election.) Only polls from September-through October were included, although for New Jersey only polls with Frank Lautenberg as the Democratic nominee were included.

In that paper, we found that actual error, as measured by the absolute value difference between survey findings and election results, was substantially higher than the reported margin of error.

While only 5% of polls should have erred by more than the margin of error, fully 25% of the 2002 Senate polls erred by more than that. Only 79% of the 2002 polls had the “correct” result and the median error was substantially higher than predicted by the z-table, which measures area under a normal curve.

In this paper we extend this analysis to a larger meta-sample of statewide presidential polls conducted in the final weeks of the 2004 campaign, and obtain strikingly different results.

### **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, in off-year elections 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. This factor would be less of a problem in a presidential election year with turnout over 50% of the voting age population.

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.<sup>1</sup> 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 impossible to estimate in advance.

While this paper and its predecessor, (Bloom 2003) are 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)

---

<sup>1</sup> 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.

Norman Ornstein put 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.)

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 usually 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 error (or *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).

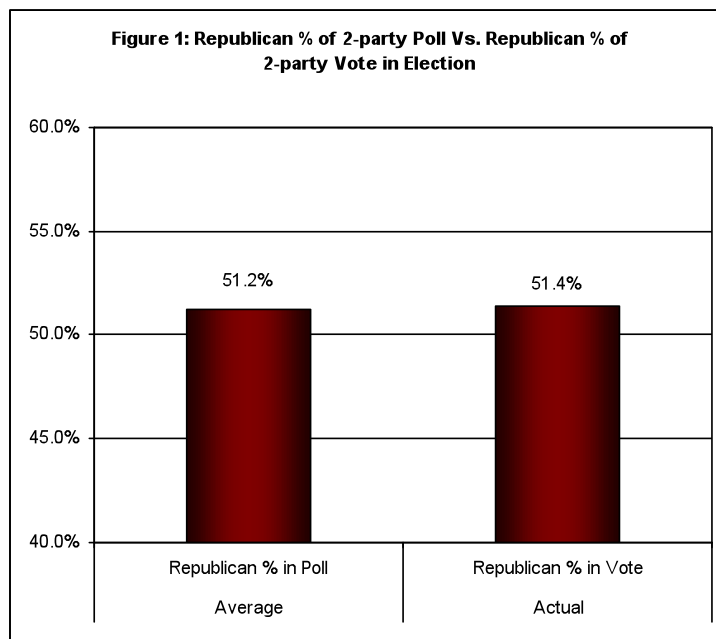
At the same time that election polls face all these challenges not faced by standard attitude polling, they also face a much stricter test not faced by other surveys– if your poll shows a certain percentage of the population favoring a tax cut no one will be able to present you with a number that is 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. As Martin, Traugott and Kennedy put it, “A peculiar position of pre-election polls is that they represent one of the few instances in which there is an external validation of the survey estimates – the actual outcome of the election.” (2003)

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. If, on the other hand, election polls are not as accurate as claimed, both public opinion professionals and consumers will need to consider that in their interpretation of election poll data.

## The Data Set

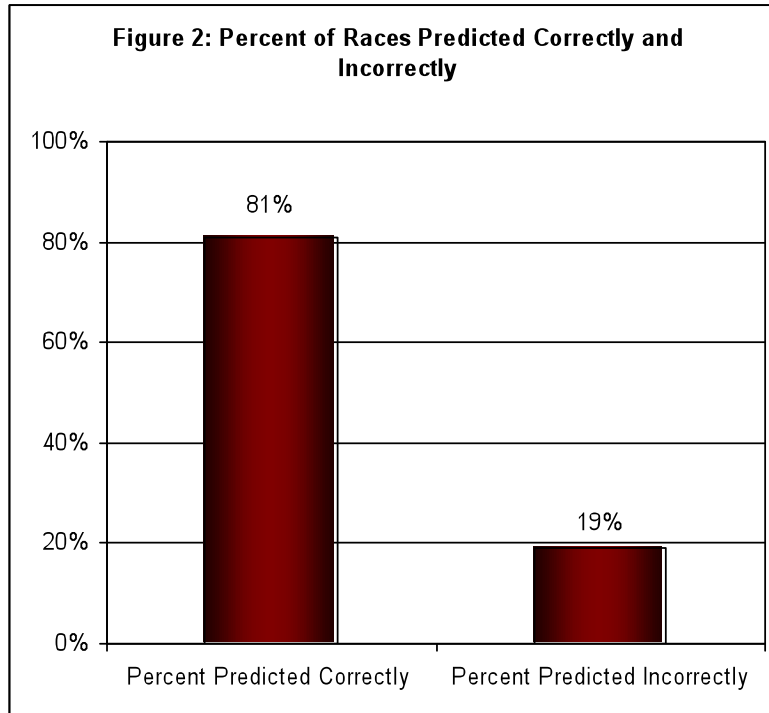
We compiled a meta-sample (a sample of samples) of 344 polls gathered from a variety of sources, including PollingReport.com and OurCampaigns.com, as well as a data set compiled by SurveyUSA.<sup>2</sup> The data set has been cross-checked, proof-read and copy-edited, but is certainly neither exhaustive nor error-free. We included all publicly available polls conducted in October with a few right at the end of September and through the beginning of November.<sup>3</sup> This represents a shorter time frame than the previous paper which included all polls beginning September 1<sup>st</sup>. The decision to shorten the time frame from the previous paper was made both because a larger number of polls were available during the presidential year, and we found that one month provided a long enough time frame to judge a real shift in public opinion versus sampling error, but not so long as to present strong possibilities of large over-time effects.

General sample parameters are shown in Figures 1-3. Bush received 51.2% of the two-party vote in the 344 polls of the sample, compared to an actual vote in the states polled of 51.4% for Bush (see Figure 1, below). Thus, the polls show no net partisan bias – an important finding, especially given the apparent bias in the exit polls. We follow Traugott (2001) in using the “Mosteller 3” (Mosteller, 1949) method of measuring survey error – looking at absolute value difference between the survey estimates and election results for major party candidates, effectively, allocating third party votes and undecideds proportionately. Given the relatively small numbers in either category in 2004, this decision has fewer consequences than it might otherwise.



<sup>2</sup> Special thanks to Tom Silver of PollingReport.com and Jay Leve of SurveyUSA for making data sets (lists of surveys, with information on dates, sample frame, and sample size) available of our analysis.

<sup>3</sup> 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 a number of instances, tracking polls with rolling (i.e. overlapping) samples were reported daily; we 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. We will be happy to share the data set with anyone in exchange for any corrections, additions or updates one might have to provide.



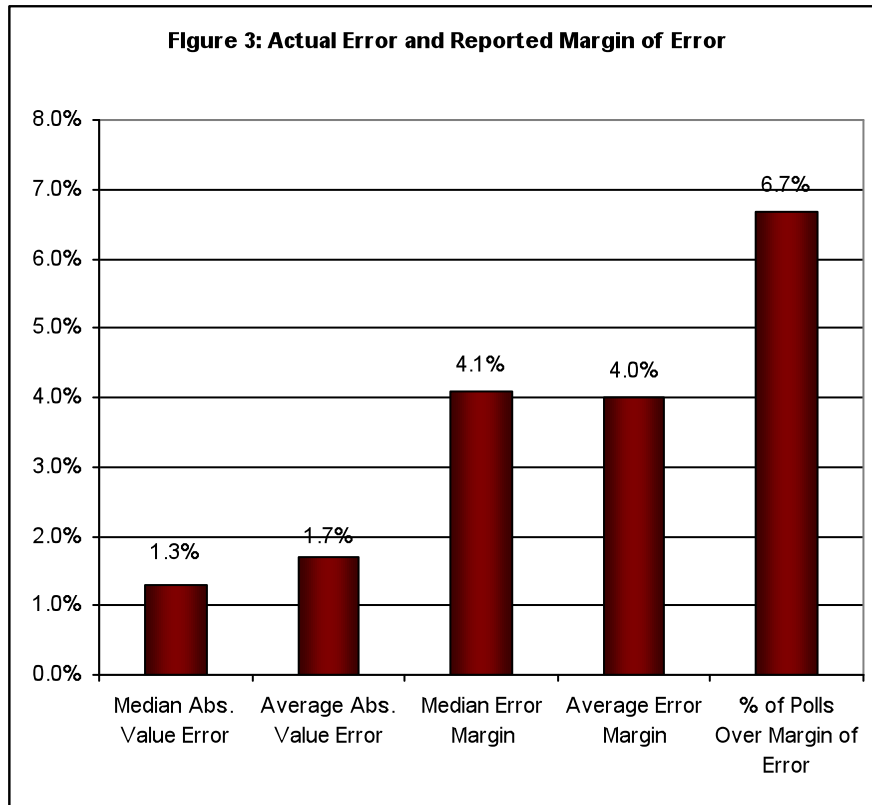
(Polls with a “tie vote” count as half a correct prediction.)

Looking at the question most important to many observers, even if it isn’t the most statistically pertinent, 81% of polls in the meta-sample predicted the winner correctly (keep in mind that states with close races are over-represented in the sample). Of course probability theory provides us with no metric against which to judge whether this represents a good, bad, or middling performance (more on this later in the paper). On average, polls reported margins of error of 3.9%, with a median error margin of 4.0%. This means that roughly 95% of the time, 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 our comparative analysis includes only major-party “decideds” we recalculated margins of error to take that into account, resulting in a slight overall increase in the effective sampling error to 4.0%, with a median error margin of 4.1% (see Figure 3, below).<sup>4</sup>

As shown in Figure 3, below, the polls in the 2004 data set averaged an absolute value 1.7% difference from the actual election results in their states, which is a substantial improvement over Senate election polls from 2002, and their average absolute value difference of 3.0%. Similarly, the 2004 polls had a median absolute value difference from state vote percentages of 1.3% (a figure the significance of which we will discuss further below), also much lower than the 2002 Senate figure of 2.1%.

---

<sup>4</sup> Thanks to Warren Mitofsky (invited address at AAPOR, 2003) and Martin, Traugott & Kennedy (2003) for this critically important observation. Because of the very small numbers of undecided voters in October, 2004, the difference is much lower than in the 2002 Senate analysis, but it is still important to take it into account in this type of analysis.



(Reported margin of error adjusted to account for reduced sample size due to omission of undecideds and minor-party voters.)

These figures are clearly well within this margin of error of around 4%, but one must recall that the margin of error 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). And in fact, as shown in Figure 3, 6.7% of cases erred by more than the margin of error. In the other measures of error the differences we have found are within the margin of error, but how do they compare with the way the polls *should have* performed as stipulated by probability theory?

Over the years, a large number of methods for testing the reliability and accuracy of election polls have been suggested (Mosteller et al., 1949; Crespi, 1988; O’Neill et al., 2002; Franklin, 2003; Martin et al., 2003). Martin et al. (2003) provide an excellent critical overview of these various measures. While Franklin (2003) and Martin et al. (2003) develop sophisticated new tools for testing the accuracy of election polls, we have presented (Bloom, 2003) far simpler and straightforward tests that we feel are the only truly appropriate measures of survey reliability. In other words, we use the tests and tools actually contained within the probability theory upon which sampling error is based, area under a normal curve, measured by the z-table. 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. The median level of error – the range within which we should find 50% of all cases in a normal distribution – is also a crucially important statistic for our analysis, as explained below.

O’Neill et al. reported 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. reported the 84% figure as if it represented an accomplishment by election polls, the reality was different. In our 2002 analysis, with its longer time-frame, we found that fully 25% of surveys fell outside of the reported margin of error (adjusted to take into account the reduced effective sample size). This was worse than O’Neill et al.’s finding, (possibly due to our inclusion of September polls and exclusion of gubernatorial races) but even their lower number is quite damning for election polls – after all, the time frame they use is short enough to rule out most large opinion shifts for all but a relatively small percentage of the included polls.

Figure 4, below, shows the distribution of error (difference between Republican percentage of two-party intent in polls and Republican percentage of the two-party Senate vote in the actual vote, rounded to the nearest percentage point) for the 2002 meta-sample, and it is not a pretty picture. That figure shows quite graphically just how far removed the actual distribution was from the expected normal distribution that year. 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 upward to account for decreased effective sample sizes.

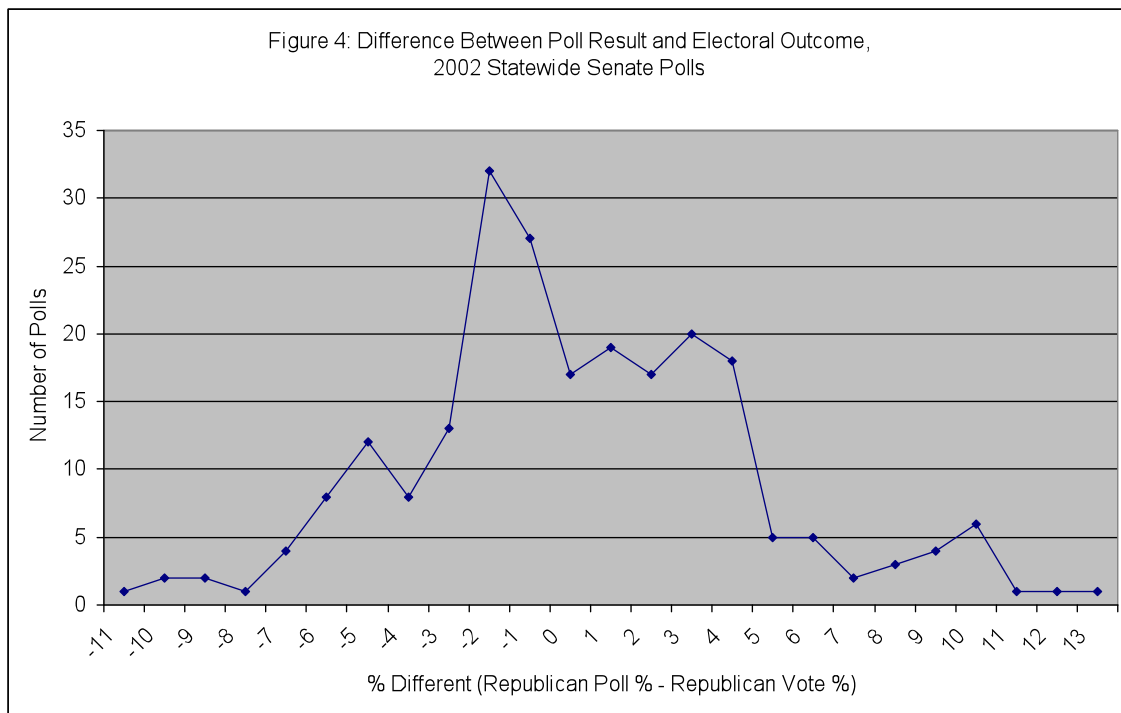
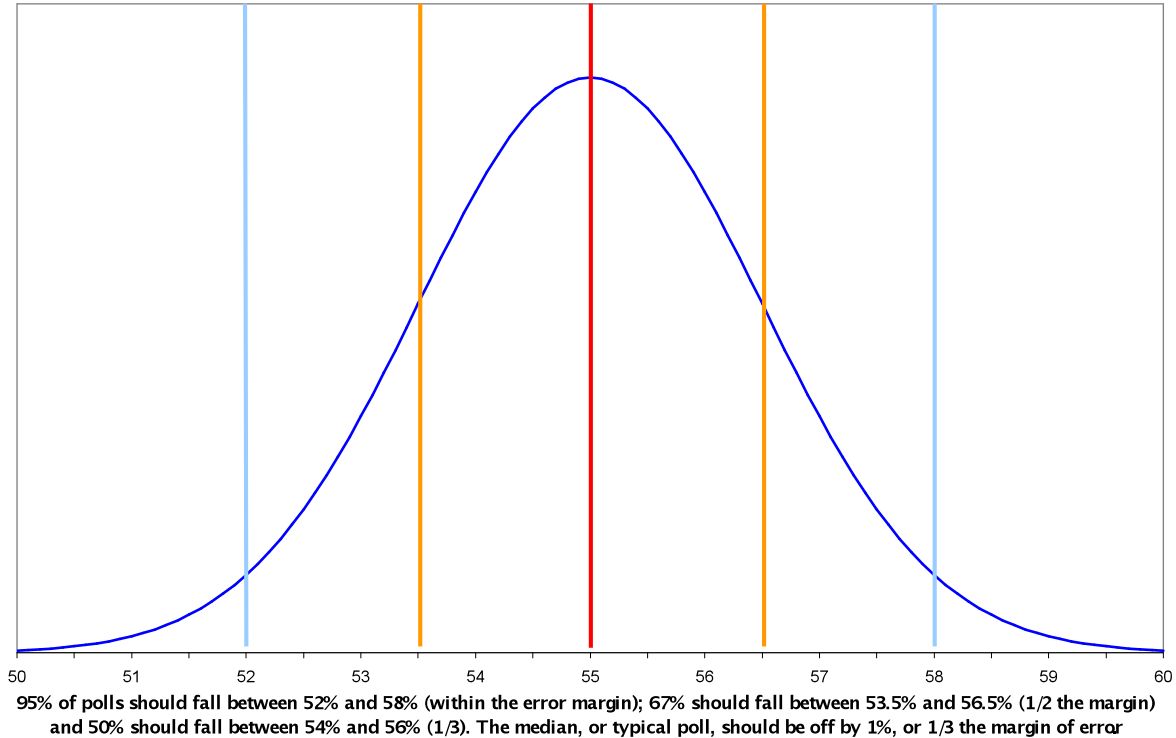


Figure 5, below, shows what the distribution should have looked like. In a hypothetical example in which a true population value for a subject of a public opinion poll is 55%, 95% of polls should be within 3% in either direction – the margin of error (the pale blue line). Two thirds should be within around one half of the error margin, in this case 1.5% (the orange lines). Finally, half the cases should fall within roughly one third of the margin of error, or 1% (line not shown). It is this standard which the 2002 Senate polls clearly failed to meet.

The z-distribution, or area underneath a normal curve, represents the same concepts, but with greater precision. In order to conduct additional tests of how the polls in our analysis compare to

predicted area under a normal curve it was 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 converted various numbers from a scale based on the median effective reported margin of error (4.354%<sup>5</sup> for the 2002 analysis) in units of 1.96. The conversion factor was thus 4.354/1.96 or 2.221.

**Figure 5: Area Under a Normal Curve.**  
**If the true population value for a survey question is 55%, results of an infinite number of surveys on that item will resemble a normal curve centered around that figure (shown with 3% margin of error).**



Using this conversion factor the next test was 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 our 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*.<sup>6</sup> The median, at 2.1%, looked a great deal better than the mean, but it was still significantly greater than the 1.5% level where it should have been. In fact, a median of 2.1% was consistent with a point in the z-distribution not of 0.67 where it should have been, but rather of 0.95. That figure was larger than the expected 0.67 by a factor of 1.41. When we multiplied this factor by the claimed average margin of error of 4.4% we got an observed total error of 6.15% (4.4% \* 1.41) – nearly two points higher than the reported margin of error.

<sup>5</sup> During this section of the paper, we report additional decimal places in order to maintain precision during these critical calculations.

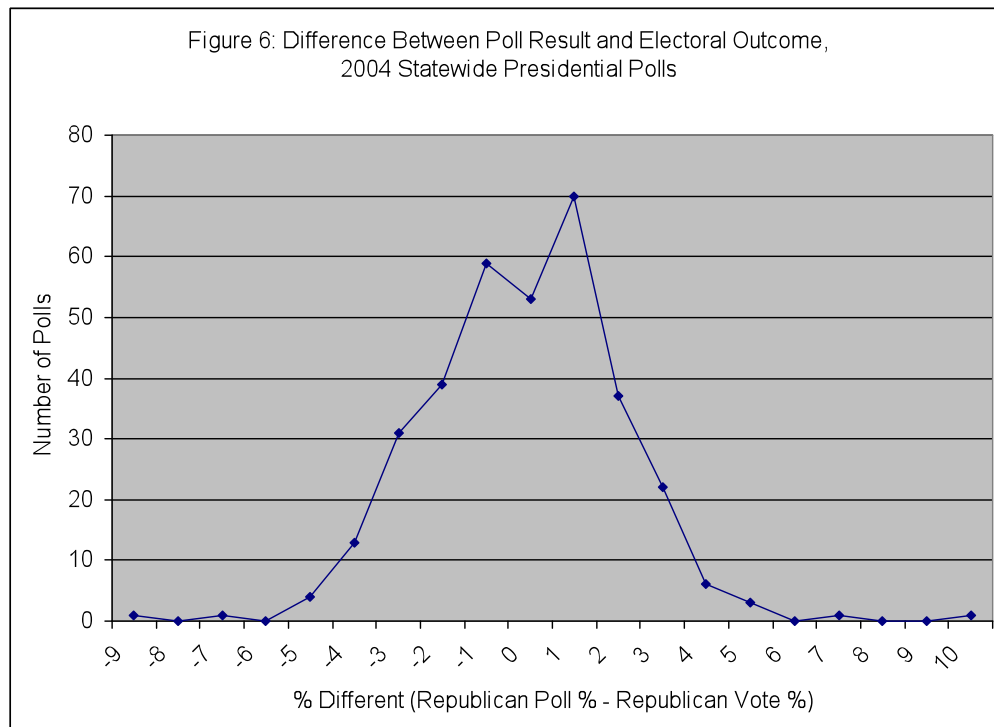
<sup>6</sup> In fact, the median is that point at which half the cases are on one side of the distribution and half are on the other, making it precisely the right comparator to the median absolute value error term used here.

Conducting the same analysis with the 2004 presidential meta-sample, we start with the median effective error margin of 4.067%, and divide it by 1.96 to obtain the z-conversion factor of 2.075%. This figure multiplied by 0.67 gets us the predicted median error of 1.390%. As mentioned above, the observed median error was just a hair below this threshold, at 1.3%, or more precisely, 1.338%. Thus, we find that unlike the 2002 Senate polls, the 2004 meta-sample of 344 statewide presidential polls performed precisely as advertised by the reported sampling error, down to a tenth of a percentage point!

Figure 6, below, shows the distribution graphically. The difference between the tightly grouped distribution of 2004 presidential polls in this chart and the haphazard and widely spaced distribution for the 2002 Senate polls is truly striking.

Another method for testing the statistical properties of this type of meta-sample of polls is to look at the tails of the distribution. As I observed earlier, 25% of the 2002 Senate polls fell outside of their reported effective error margins, meaning that 75% lay 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 suggested a true margin of error of 7.42% ( $4.4\% * 1.70$ ) – even higher than the test suggested by the distribution’s median.

Again, the 2004 presidential polls performed a great deal better than the 2002 Senate polls. In this meta-sample, 93.3% of surveys fell within their margins of error, while 6.7% (23 surveys) differed from the election results by more than the margin of error. While this number is somewhat higher than it should be, when we keep in mind that this is only a difference of six polls from the 17 that probability theory would have predicted, the difference is not substantively large. Thus, by this measure as well, the 2004 presidential polls performed very well. Using the z-table again, the point in the z-distribution outside of which 6.7% of the cases should be found in a 2-tailed test is 1.83 rather than the 1.96 for 5% of cases. Since 1.96 is larger than 1.83 by a factor of 1.07, this suggests a true margin of error of  $1.07 * 4.07$  or 4.35. Thus, based on this measure, the 2004 statewide presidential polls did not quite measure up to par, but they were very close.



### **Assessing the Accuracy of Individual Pollsters**

When 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 SurveyUSA for its controversial use of automated telephone interviewing systems. Since we had 30 or more polls from each of these firms in our 2002 data sets, we were able to assess the reliability of these two firms’ surveys overall, finding that despite a few late missed calls, Zogby did fairly well overall, and SurveyUSA actually out-performed the field by some measures. In the 2004 analysis we have another firm using automated calling technology – Scott Rasmussen. If we find that both SurveyUSA and Rasmussen perform at similar levels to the rest of the field, we should lay to rest the question of whether their new methodology should be considered a scientifically legitimate alternative to conventional telephone surveys, even if some of us are uncomfortable with it.

In the 2004 data set, we have nine polling operations with at least 10 surveys, so a broader look at some of the field’s more prolific pollsters is possible. The findings are displayed in Table 1, below. Looking first at the question of whether the polling operation had a net partisan bias, the range was from around a 1% net Republican bias (Quinnipiac University and Strategic Vision, the only partisan pollster in the group) to a 1% net Democratic bias (Rasmussen and Research 2000). Almost all the polling operations in this group outperformed the overall meta-sample with regard to the median absolute value difference between the poll and the electoral outcome – only Gallup and Zogby had higher error than the overall sample, both substantially so. Looking at the percentage of polls with error beyond the margin of error, Gallup once again stands out as having particularly large instances, with four out of 15 polls showing error higher than the survey’s reported error margin. Looking at predictive accuracy, Gallup again ran into trouble, with just over half of their surveys correctly predicting state outcomes. ARG was even worse in this category, with incorrect findings in over half of its surveys, despite its very low levels of error. Because different pollsters worked in different states, however, this number can be misleading.

As a group, these prolific pollsters averaged 29 surveys each and accounted for 227 of the sample’s 344 surveys. The remaining pollsters averaged only 2 surveys and totaled 117 surveys. The more prolific pollsters outperformed the less experienced pollsters by every measure. The prolific pollsters had lower partisan bias (-0.1% vs. -0.3%), lower absolute value error (median of 1.2% vs. 1.7%), fewer cases outside the margin of error (4% vs. 12%) and slightly higher predictive accuracy (82% vs. 79%).

Not only did pollsters with a lot of election poll experience perform better than their less experienced peers; surveys conducted in states with a large number of other surveys were more reliable than surveys conducted in states with few other surveys, although the differences are not as large. As seen in Table 1, in states with fewer than ten surveys in the sample, both net partisan bias and error levels were higher than in states with more surveys. More surveys were outside the margin of error in states with fewer polls (9% vs. 5%) but predictions were more accurate, apparently due to the fact that this group included more states that were not as closely contested.

Not surprisingly, the number of polls in a state was closely related to the margin of victory of the winning candidate. Splitting the sample based on whether a state was a closely contested “battleground state” – operationalized here simply as a state with a margin of less than 8% -- produced a nearly identical breakdown to the one based on the number of polls per state. By every measure except predictive accuracy, error was lower in the battleground states than in the

states won by large margins. The “blowout” states averaged only 5 polls each, while the “battleground” states averaged 19 polls. Obviously, in the blowout states one can be off by several points and still predict the outcome correctly.

**Table 1. Comparisons of Poll Numbers vs. Actual Votes, by Pollster and Other Factors.**

	# 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**
ARG	16	49.4 %	49.8%	-0.4%	4.1%	1.0%	0.7%	0%	44%
Gallup	15	50.6%	50.2%	0.4%	3.4%	2.2%	1.8%	27%	53%
Mason Dixon	36	52.2%	52.0%	0.2%	4.0%	1.3%	0.9%	0%	93%
Quinnipiac U	13	49.7%	48.6%	1.1%	3.3%	1.5%	1.3%	8%	85%
Rasmussen	29	51.0%	52.1%	-0.9%	4.4%	1.6%	1.0%	7%	98%
Research2000	15	48.3%	49.2%	-1.1%	4.1%	1.5%	1.2%	0%	93%
Strat Vision	36	50.8%	49.8%	1.0%	3.6%	1.3%	1.0%	0%	74%
SurveyUSA	30	52.1%	52.5%	-0.4%	3.9%	1.5%	1.1%	0%	97%
Zogby	37	50.1%	50.7%	-0.6%	4.2%	1.9%	1.7%	5%	77%
Prolific	227	50.8%	50.9%	-0.1%	3.9%	1.5%	1.2%	4%	82%
Non-Prolific	117	52.1	52.4	-0.3%	4.1%	2.1%	1.7%	12%	79%
< 10 in State	130	53.7%	54.0%	-0.3%	4.1%	1.9%	1.4%	9%	98%
10+ in State	214	49.7%	49.8%	-0.1%	3.9%	1.6%	1.3%	5%	71%
> 8% Margin	108	54.4%	54.8%	-0.4%	4.2%	2.1%	1.6%	11%	98%
Battleground	236	49.8%	49.8%	-0.0%	3.9%	1.6%	1.3%	5%	73%
Week 1	29	50.6%	49.9%	0.7%	4.0%	2.2%	1.8%	14%	74%
Week 2	21	51.9%	52.0%	-0.1%	4.4%	2.5%	2.5%	5%	79%
Week 3	70	52.1%	52.3%	-0.3%	4.0%	1.7%	1.3%	10%	80%
Week 4	99	51.4%	51.6%	-0.2%	4.0%	1.7%	1.5%	5%	84%
Week 5	125	50.6%	50.9%	-0.3%	4.0%	1.5%	1.2%	5%	81%
Total	344	51.2%	51.4%	-0.2%	4.0%	1.7%	1.4%	7%	81%

\*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.

Table 2 shows Pearson’s correlations of some of these factors on a variety of measures of survey reliability. Larger numbers of polls in a state are indeed associated with lower error levels, but also with more missed predictions (due to the fact that states with closer races have more polls). Larger numbers of polls by a pollster are even more strongly associated with lower error. Finally, states with larger margins of victory also have higher survey error, although they also have more accurate predictions of winners and losers.

**Table 2: Pearson’s Correlations on Survey Error.**

	# of Polls in State	# of Polls by Pollster	Size of Vote Margin
Absolute Value Error	-.131* (.015)	-.181** (.001)	.193** (.000)
Over the Error Margin	-.051 (.348)	-.177** (.001)	.119* (.027)
Predicted Correctly	-.322** (.000)	.089 (.100)	.320** (.000)

\* Statistically significant at .05 \*\*Statistically significant at .01 (both two-tailed tests)

## Multivariate Analysis

To test whether the differences among categories of pollsters and states hold up when other important factors are held constant, we conducted a simple OLS regression analysis with poll error as the dependent variable; the results appear in Table 3, below. Specific hypotheses tested relate to the impact on survey error of several factors: (1) number of polls in the state; (2) number of polls by the individual pollster; (3) the polls' margins of error; (4) the end date of the poll; and (5) how close the actual vote was in the state.

(1) Hypothesis 1: Surveys in states with more surveys will have lower error than those in states with fewer surveys. As much as pollsters like to believe that our own methods are the best combination of best practices, everyone is susceptible to peer pressure. To put it bluntly, if pollster x does a survey in Wyoming, where it is the only poll, she is going to release the results no matter what they say. But if she is polling in Wisconsin, she's almost certainly comparing notes with other recently released polls. If her results come out very different from the others, she's likely at the very least to take a second and third look at her data set and weighting procedures for any errors. If that ends up producing results closer to the other polls, all the better – that's where they should have been anyway. If the results are way off, she might choose simply not to release them, an option she wouldn't have even considered without the other polls for comparison.

(2) Hypothesis 2: Surveys conducted by pollsters with a large number of surveys in the sample will have lower error than those conducted by pollsters with smaller numbers. Election polling is different than conventional surveying behaviors and attitudes, partly because the stakes are higher for being right or wrong. As a result, experienced election pollsters almost always weight their sample to known or projected population parameters. If done well, this sort of weighting should literally reduce sampling error in that it is sampling error (along with nonresponse bias) that can lead to the over- or under-representation of specific groups in the sample. (The fact that our overall sample numbers show actual error behaving very nearly precisely the way sampling error should may thus actually be a product of lower effective sampling error offset by some of the other types of error discussed at the beginning of the paper.) Less experienced pollsters might not weight at all, or might do so more haphazardly.

(3) Hypothesis 3: Surveys with lower margins of error should have lower actual error. This hypothesis speaks for itself – lower margins of error resulting from larger sample sizes should produce more reliable survey estimates.

(4) Hypothesis 4: Surveys in the field closer to Election Day should have lower errors than those conducted earlier. One of the additional possible sources of error (in addition to sampling error) is that there might be real shifts in the electorate's voting preferences over time. As a result, it is reasonable to think that polls conducted closer to Election Day should have lower error levels than those done earlier in the campaign.

(5) Hypothesis 5: States with closer races should have lower error rates than states won by large margins. This hypothesis is closely related to Hypothesis 2 – in closely-fought “battleground” states the stakes are higher. Not only are there more polls (again, see Hypothesis 2), but the closer the race, the closer the level of attention pollsters are likely to pay to the details of their data set and their methodological practices. This higher level of attention should be associated with lower error levels.

Dummy variables for the most prolific pollsters are included to test whether the differences observed in Table 1 hold up when other factors are held constant. Other variables designed to test the hypotheses listed above include the number of polls in a state, the number of polls by the pollster, the margin of error, the end date of the poll (larger numbers are closer to election day), and finally the winner's vote margin.

Of all the variables included in the model, only the end date of the poll is statistically significant. The closer to Election Day, the smaller is the difference between the survey and the elections results. Thus, only one of the hypotheses listed above beats its "null hypothesis" that there will be no statistically significant difference.

Looking at the overall model statistics, the  $R^2$  is only .132 (adjusted  $R^2 = .095$ ), indicating that only around 10% of the variance in survey error is caused by factors 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 margin of error itself). While that is not the case (the F-statistic for the model is statistically significant) the large majority of the survey error certainly is explained by factors not included in the model, which we believe to be random error as predicted by probability theory. (In contrast the R-square for the 2002 Senate surveys was .326 and a number of coefficients were statistically significant.)

**Table 3. Multivariate Regression Analysis of Survey Error.**

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	4301.064	1507.104		2.854	.005
# of Polls in state	-.010	.012	-.062	-.862	.389
# of Polls from pollster	.033	.068	.351	.481	.631
ARG	-1.329	1.020	-.204	-1.303	.194
Zogby	-.964	2.336	-.218	-.413	.680
Rassmus	-1.278	1.870	-.259	-.684	.495
Quinnipiac	-.589	.865	-.082	-.681	.496
MasonDixon	-1.713	2.345	-.383	-.731	.465
Gallup	.172	.860	.026	.200	.842
SurveyUSA	-1.309	1.955	-.270	-.670	.504
Research 2000	-.951	.969	-.142	-.981	.327
Strategic Visions	-1.504	2.347	-.336	-.641	.522
modified error margin	.181	.157	.072	1.150	.251
poll end date	.000	.000	-.167	-2.853	.005
2 party Vote Margin	.021	.012	.124	1.729	.085

Dependent Variable: absolute value of Error

$R^2 = .132$ , Adjusted  $R^2 = .095$ , SEE = 1.305, F=3.586 (P=.000)

**Table 4. Multivariate Regression Analysis of Survey Error**

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	3953.024	1445.188		2.735	.007
# of Polls in state	-.003	.011	-.019	-.277	.782
# of Polls from pollster	-.009	.005	-.098	-1.747	.082
modified error margin	.179	.136	.071	1.314	.190
poll end date	.000	.000	-.153	-2.735	.007
Absolute Value 2 party Vote Margin	.023	.012	.134	1.937	.054

Dependent Variable: absolute value of Error

$R^2 = .088$ , Adjusted  $R^2 = .074$ ,  $SEE = 1.320$ ,  $F=6.497$  ( $P=.000$ )

Since there is a great deal of overlap between the pollster dummy variables and the number of polls from pollster variable, we also ran the regression without the pollster variables. The results in Table 3 show a reduced R-square and increased standard error of the estimator. Again, the only variable with a statistically significant coefficient is the poll end date, although this time the number of polls variable comes fairly close at .08 significance. Overall, the story of the reduced model is very much the same as the one that included the pollster variables – the vast majority of variance is explained by variables outside the model and only the proximity of the survey to the election makes a significant difference. It is our belief that the unexplained variance is evidence that sampling error really is randomly distributed across polls as it is supposed to be.

### Conclusion

This meta-analysis of survey data shows that the presidential election polls were far more reliable than the 2002 Senate polls. In terms of median error, the 2004 statewide presidential polls performed precisely as advertised by the reported sampling error, down to a tenth of a percentage point. In terms of percentage with errors larger than the margins of error, 6.7% (23 surveys) are in this category, with more experienced pollsters providing more accurate and reliable poll results than the less experienced groups. Overall, these results show a vast improvement over the 2002 analysis where 25% were outside their reported margins of error. Another important finding is that polls also showed no partisan bias. Regression analysis shows that, all else held constant, only the end date of the poll was significantly related to survey error (Sig.=.005).

Much more remains to be done here, including expanding the data set and analysis to include Senate polls from 2004. In particular, this would allow us to determine whether there was something about polling in an off-year election that adversely affected reliability, or whether Senate races might be just more difficult to achieve the levels of reliability found in presidential polling due to the fact that Americans don't have as strong feelings about their Senate vote as they do their presidential vote. We will also be applying more sophisticated analytical techniques including time series analysis to attempt to take into account real shifts in voting intent over time.

While Martin et al. (2003) and Franklin (2003) contribute a great deal with their sophisticated multivariate methodologies (which will doubtless prove very useful, especially with regard to predictive accuracy), we believe that the analysis presented here is in many ways more appropriate and more broadly applicable in the public discourse about poll reliability. After all, pollsters make no claims about the natural logarithm of the odds ratio, but they do make claims about what percentage of cases should fall within particular distances of the election outcome.

At the very least, election polls should report a margin of error based on the number of “decideds” in their sample rather than the entire sample size. Even then, however, the actual error level is evidently sometimes substantially higher than claimed. On the other hand, in some ways the margin of error term, representing the tails of the distribution, presents a misleadingly bleak picture of the likely error. One possible solution would be to report the median expected divergence from the election results (in this case 1.5%) or the actual median divergence from previous elections (in this case 2.1%) in addition to the margin of error. This would give the poll consumer a better idea of how much polls *typically* diverge from election results. The question still remains, “reliable compared to what?” but we 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.
- Bloom, Joel D. 2003. "Reliable Compared to What? Empirical Tests of the Accuracy of Election Polls, 2002." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, May 16-19 2003, in Nashville, TN.
- Bloom, Joel D. 2001. *Intentioned Voters, Unintended Outcomes: Low Information Rationality and Split-Ticket Voting in the United States*. A doctoral dissertation in political science, University of Michigan.
- 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.
- Huffington, Ariana. November 14, 2002. "The Pollsters Can't Hear The Silent Majority" (<http://www.ariannaonline.com/columns/files/111402.html>)
- 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](http://elections.gmu.edu/VAP_VEP.htm))
- Morin, Richard. "Smackdown in Maryland: RBS versus RDD." 2003. *Public Perspective* 14, Number 1; 7-9, 41.
- Mosteller, F., Hyman, H., McCarthy, P., Marks, E., and Truman D. 1949. *The Pre-Election Polls of 1948: Report to the Committee on Analysis of Pre-Election Polls and Forecasts*. New York: Social Science Research Council.
- 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 (NCP) 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. 2001. "Assessing Poll Performance in the 2000 Campaign." *Public Opinion Quarterly* 65:389-419.
- Traugott, Michael W. and Paul J. Lavrakas. 2000. *The Voter's Guide to Election Polls, 2<sup>nd</sup> Ed.* New York: Chatham House.