A Review and Proposal for a New Measure of Poll Accuracy

Elizabeth A. Martin (U. S. Census Bureau), Michael W. Traugott (University of Michigan), and Courtney Kennedy (University of Michigan)

Introduction[1]

In the aftermath of the polling industry's disaster in the 1948 election, a distinguished group of social scientists and statisticians quickly mounted an intensive review of the election polling procedures and results to evaluate what had gone wrong. In just five weeks, they produced a remarkably thorough assessment, published by the Social Science Research Council (Mosteller et al., 1949). Mosteller's chapter on "Measuring the Error" was a lucid review of the major poll estimates, and his work established the measures that have been used ever since to evaluate the accuracy of election polls.

The authors of the SSRC report considered their work to be preliminary, and expressed the hope and expectation that "definitive and more leisurely studies" of the problems they identified would be conducted. Since then, the merits of Mosteller's error measures have been discussed (by e.g., Mitofsky, 1998); but curiously, there has been relatively little follow-up work by statisticians to improve or evaluate them. In the ensuing years, there have been occasional controversies about the accuracy of pre-election polls, including the underestimation of Ronald Reagan's victory in 1980 and the overestimation of Bill Clinton's victory in 1996. From time to time, there have also been more general calls for a review of the accuracy of election polls. For example, in 1984 the Panel on Survey Measurement of Subjective Phenomena recommended establishing "a panel or committee to evaluate the performance and methodology of election polls," and noted that "a regular review of the accuracy of such forecasts could be of use both to the survey industry and to the public" (Turner and Martin 1984:314). Since the advent of the modern polling period, the role of pre-election polls in forming the image of the entire industry has grown because, unlike most surveys, pre-election forecasts may be judged against an external criterion of validity – the actual outcome of an election. The performance of pre-election polls in forecasting elections may shape public perceptions of the accuracy of surveys more generally.

Across this same period, political strategists and social critics from all domains of the political spectrum have challenged the accuracy of polls and the role they play in contemporary society. Polls in recent elections have been charged with partisan bias (by, e.g., Huffington, 1996, 1998, and 2001; Ladd, 1996). Claims of bias should be addressed empirically in order to evaluate them. This argues for a regular, independent review of the polls, in the good years as well as the bad. We also think it is time to take a fresh look at measures of election poll accuracy. Advances in statistical theory and estimation of error since 1949 might yield alternative and perhaps better measures of election poll accuracy. In this paper, we offer some preliminary suggestions, and we look forward to their review by others with an eye toward improving them.

Background and Data

After the 1948 election, Mosteller proposed eight measures for evaluating the accuracy of election forecasts, six of them based upon the estimated proportion of the vote that a (leading) candidate received or the difference in the estimated margin between the leaders.[2] In 1998, Mitofsky noted the lack of consensus about the best measure for gauging poll accuracy and compared results for four of Mosteller's original methods. He decided that the best choice was between Mosteller's Measures 3 (average deviations for each party or candidate) and 5 (the difference in the differences between the leading candidates in the polls and the actual results); he favored Measure 5. Measure 3 captures "the error by averaging the deviations in percentage points between predicted and observed results for each party (without regard to sign)," and Measure 5 uses "the difference of the oriented differences between predicted and actual results for the two major candidates" (Mosteller 1949:55; one might quibble with Mosteller's lack of formulas to define his measures). Measure 3 corresponds to the error on the candidates, and Measure 5 to the error on the margin between the two leading candidates. When there are just two candidates, Measure 3 is half of Measure 5 if there are no undecided voters. These measures have been used in subsequent evaluations of election poll accuracy, such as Traugott's (2001) evaluation of poll performance in the 2000 campaign and the National Council on Public Poll's (NCPP) (2002) review of the 2002 election polls.

As Mitofsky (1998) noted, handling undecided voters is a significant problem that was not addressed by the SSRC report. Most of the methods defined by Mosteller rely on percentage point differences, and hence are affected by the size of the undecided category (and, for Measure 3, the size of any third party or other parties' candidate's share). Some polls allocate undecided voters and some do not, and

[1] This paper reports the results of research and analysis conducted collaboratively by Census Bureau staff and researchers from the University of Michigan. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or the University of Michigan. We thank Bob Fay for assistance on calculation of the variances, and Patrick Cantwell, Diane Colasanto, Bob Fay, Warren Mitofsky, Colm O'Muircheartaigh, Paul Siegel, and Eric Slud for helpful comments on an earlier draft.

[2] An additional measure was a chi-square test, deemed too burdensome to calculate 50 years ago, before the advent of modern computing. Another was based on projections of electoral votes, a practice common in the 1940's that disappeared with the advent of telephone surveys and national samples that did not represent individual states. Changes in the cost and technology of polling and the use of sophisticated statistical modeling techniques brought this practice back in the 2000 presidential election (Traugott, 2001). It may become more prevalent in the future.

measures that rely on percentage point differences (or differences of differences) will not be comparable between such polls. However, allocating the undecided requires assumptions (which may or may not be supportable) and/or additional information (which may not be available). The 2002 NCPP review did not adjust its error calculations for the size of the undecided, "in order to avoid an arbitrary decision about how to allocate" them.

The Mosteller measures focus on the closeness of estimates to an actual election outcome. Another issue is bias, or the extent to which polls over- or under-estimate a given party's share of the vote. This has been a contentious element of the statistics used to measure accuracy because sampling error is assumed to be symmetrical around a poll-based estimate. Measure 3 does not permit analysis of bias, because it is the average of (unsigned) deviations between predicted and observed results for a candidate. Mitofsky used a crude version of Measure 5 to address Ladd's 1996 claim that the "election polls have frequently over-estimated the Democrats' share of the vote" by counting the number of polls that overstated Democratic or Republican strength. He concluded that on this point Ladd was correct, since more than twice as many polls overstated as understated the Democratic share of the vote. But most evaluations of poll accuracy examine absolute errors (e.g., Traugott, 2001; NCPP, 2002) and hence do not examine bias. Bias is also a difficult issue to deal with because one component of the difference between a poll estimate and a candidate's actual standing in the election can be attributed to sampling error, a random statistical element of the study design. Any assessments of bias have to account for differences in estimates that cannot be explained by chance alone.

This paper proposes a new measure of the *predictive accuracy* of election polls that permits examination of both accuracy and bias, and applies it to summarize results from a number of pre-election estimates. We first briefly review past measures of accuracy, then describe the new measure. We apply it to three prior presidential elections (1948, 1996, and 2000) and compare its results with Mosteller's measures. Then the new measure is applied to the results of 548 state polls from gubernatorial and Senatorial races in the 2002 elections[3] to examine the accuracy of pre-election polls in "off year" races. These polls are often conducted by smaller, less experienced firms than those that conduct the major pre-election polls in presidential elections, and they often include polls conducted for candidates that are released to the media.

The analysis also includes polls conducted early as well as late in the campaign, permitting an assessment of whether the "accuracy" of polls measured in relation to the eventual outcome of the election changed as Election Day approaches. This compilation of polls is different from the 159 polls analyzed by NCPP, which excluded partisan polls and polls that were released too far in advance of Election Day. We first address the elementary question of whether the polls supported statistical inferences about election outcomes. We apply the proposed new measure of accuracy to examine the potential biases of the polls, and consider some possible sources of those biases we discover.

Results

1. A Review of Past Measures of Poll Accuracy

The systematic evaluation of polling accuracy begins with the report by Mosteller et al. (1949) after the debacle of 1948. The Mosteller team acknowledged a number of problems associated with producing estimates of election outcomes from pre-election polls. This is always easier when there are only two candidates in a race rather than three or more. It also gets more complicated conceptually when one considers the "total error" in a survey rather than the difference between the outcome and the estimate for a single candidate. In presidential elections, the number of third-party candidacies receiving more than 5% of the vote has remained important, although third-party candidates are less likely to appear in other statewide offices. The timing of the projection relative to Election Day can also present problems. Campaigns matter, and last minute shifts can occur in the electorate.

Crespi (1988) conducted the only other major study of the accuracy of pre-election polling. He assembled 430 final pre-election polls that had been publicly available or disseminated after 1979. Almost three-quarters were for races other than President, and more than 400 were for sub-national geographical units, mostly states and municipalities. Crespi calculated the percentages favoring each candidate after excluding the undecideds, and then considered three different measures of the deviation from the election result: the difference in the outcome for the winning candidate, the mean percentage difference in the outcome for the top three candidates, and the largest difference between the poll result and the actual outcome for any of the top three candidates. The three measures were highly correlated (between .81 and .93), and Crespi used the first measure because it was simplest to calculate. Using an ordinal measure of the length of time the interviews were conducted before the election, Crespi found that accuracy increased in polls taken closer to an election ($r = .21$).

Rademacher and Smith (2001) looked at 79 state-level estimates of presidential races in 2000. Their analysis paralleled an NCPP analysis of the national polls, using the same measure of "candidate error" – taking one-half of the absolute difference between the top two candidates in the poll and the difference between their electoral results in the state. This approximates Mosteller's Measure 5, although it

[3] We gathered all the state 2002 polls we could locate, including polls from nationaljournal.com, The Hotline, harrisinteractive.com, the 2002 NCPP report, an ABC News file provided by J. Krosnick, and http://www.dcpoliticalreport.com/2002/polls02short.htm. We included partisan as well as non-partisan polls. All polls were fielded on or after Labor Day. The variables analyzed were coded based on publicly reported information, which was not available for all polls.

ignores the relative standing between the two candidates. And the measurement in absolute terms eliminates the possibility of investigating systematic errors in the estimates. As a result, both Rademacher and Smith and the NCPP report also looked at whether the polls predicted the correct winner. Using these dual criteria, Rademacher and Smith found that state polls did not compare favorably with national polls. The "candidate error" was about 70% greater (averaging 1.9 percentage points compared to 1.1 for the national polls), and in about one in five cases was greater than sampling error would suggest. In 15% of cases, the polls suggested the wrong candidate would win, although many of these estimates were in states that turned out to be very close.

The NCPP reviewed 159 published polls conducted during the last two weeks of the campaign for gubernatorial and Senate elections in 2002. The average candidate error was 2.4 percentage points, and 84% of polls differed from the election outcome by less than their margin of error. 14% of polls predicted the wrong winner (a figure very close to Rademacher and Smith's finding for state polls in 2000).

Summarizing this prior work, a first reaction is how limited systematic study has been. Secondly, most of the measures of accuracy have focused on the relationship between estimates for a single candidate or party's vote and the outcome of the election or on the difference between the two leading candidates. These assessments have predominantly been conducted on the basis of the absolute value of these measures, eliminating the possibility of evaluating bias. While there has been some discussion of the utility of considering total error, this has been a difficult concept to operationalize. We address these issues in proposing a new measure of polling accuracy and assessing its utility.

2. A New Measure of Predictive Accuracy

To examine how accurate and unbiased pre-election polls are, we introduce a measure based upon the following odds:

1. The odds on a Republican choice in a given poll, defined as r_i/d_i, where r_i is the proportion of respondents favoring the Republican candidate and d_i is the proportion favoring the Democratic candidate in the *i*th poll, and where $n_i = r_i + d_i$, or the total number of respondents who favor Democrats and Republicans in the poll.

The odds has a clear interpretation: odds greater than 1.0 imply a Republican lead in poll *i*, odds less than 1.0 imply a Democratic lead, and odds of 1.0 imply a tie.

A poll conducted for the 2002 Alabama governor's race is illustrative. A total of 900 people were interviewed, with 39% favoring the Democratic candidate, 45% the Republican, and 16% undecided. We ignore the undecided, and form the odds $r_i/d_i = .5357/.4643 = 1.154$. Note that the effective sample size n_i is reduced to 756, not 900, for this measure. Note also that the same value of the ratio is obtained using numbers or proportions, regardless of whether the undecideds are included or excluded from the denominator ($405/351 = .5357/.4643 = .45/.39 = 1.154$).

2. The odds on a Republican choice in an actual election, defined as R_{jk}/D_{jk}, where R_{jk} is the number (or percentage) of voters who favor the Republican candidate and D_{jk} is the number (or percentage) of voters who favor the Democratic candidate in an election for the *j*th office in the *k*th state.

In the 2002 Alabama governor's race, the Republican won a cliffhanger with 50.1% of the vote, or 672,225 votes to 669,105 for the Democratic candidate. Thus, the election odds is 1.005–very close to a tie, but greater than 1.0, indicating a Republican victory.

We calculate an odds ratio by dividing the odds for poll *i,* office *j*, in state *k,* by the election odds:

$$\text{odds ratio}_{ijk} = (r_{ijk}/d_{ijk}) \,/\, (R_{jk}/D_{jk})$$

The odds ratio also has a clear conceptual interpretation: an odds ratio of 1.0 implies the poll and the election odds are in perfect agreement, with exactly the same *relative* distribution of voter preferences between the top two (Republican and Democratic) candidates. The farther from 1.0 an odds ratio is, the worse the poll performed at predicting relative preferences in the election. An odds ratio less than 1.0 implies the poll favored the Democrat compared to the actual election result, while an odds ratio greater than 1.0 implies the poll favored the Republican compared to the election result. Some departures from 1.0 are to be expected due to sampling error, of course. Departures that exceed sampling error can be regarded as measures of the bias characterizing polls.

We transform the odds ratio by taking its natural log to make it symmetric and to simplify the calculation of the variance[4]. Thus, we define our measure of predictive accuracy A as

$$A_{ijk} = \log [(r_{ijk}/d_{ijk}) \,/\, (R_{jk}/D_{jk})] \qquad (1)$$

$$\text{Variance } (A_{ijk}) = 1/n_i r_{ijk} d_{ijk} \qquad (2)$$

A may take on values of zero, or positive or negative values.

- a value of 0 reflects perfect agreement between a poll and election result (A is zero when the odds ratio defined above is 1.0).
- a significantly negative value indicates a poll is biased in a Democratic direction (that is, its distribution was too Democratic compared to the election outcome).
- a significantly positive value indicates a Republican bias.
- negative magnitudes are comparable to positive (unlike the odds ratio).

[4] We are grateful to Bob Fay for deriving the formula for the approximation to the variance.

Again illustrating with the poll conducted for the 2002 Alabama governor's race: with odds of 1.154 on a Republican choice in the poll and 1.005 in the election itself, the odds ratio is 1.154/1.005=1.148 and the log of the odds ratio, A, is .138. A positive value of A indicates the poll overstated preferences for the Republican candidate, compared to the election outcome. Was it biased?

To assess its bias, we construct a confidence interval around zero, the expected value of A in the absence of bias. The variance of A is $1/nrd$ or $1/(756*.5357*.4643) = .005$ with standard error .073, so a 95% confidence interval includes 0 ± .143. Since the value of A is within the confidence interval, we conclude the poll is not significantly biased.

This measure has several advantages compared to traditional measures that rely on percentage point differences to measure a discrepancy. First, the odds and log of the odds ratio are amenable to multivariate analysis and modeling using log linear methods. That is to say, they can become dependent variables in equations where the explanatory factors can be either methodological attributes of the pre-election polls, such as timing or sample selection procedures, or contextual factors that distinguish elections, such as type of race, state attributes, or incumbency.

Second, by excluding them, the measure circumvents the problem of allocating prospective voters who are undecided. Indeed, the odds is the natural way of representing what seems to be fairly well established in public opinion measurement, which is that the *relative* proportions in substantive categories are often unaffected by changes in the size of the no opinion category. This conclusion is suggested by experimental research showing that the presence or absence of an explicit no opinion or middle category (cf. Schuman and Presser, 1979; Presser and Schuman, 1980) affects the percentage providing each substantive response, but not the relative proportions. More recent research suggests that question form and order can also affect the proportion of undecided in a pre-election poll (McDermott and Frankovic, 2003).

In effect, we assume a distribution of "undecideds" in proportion to the actual distribution of candidate preferences in the polls. Other assumptions are possible. As a general assumption, this is probably an oversimplification. Research shows that proportional allocation may provide a poorer prediction of the voting behavior of undecided voters than other allocation strategies (see, e.g., Visser et al. 1996). However, this simplifying assumption is useful for our purpose of examining the performance of a large number of polls, for which we lack information that might be used to devise some other allocation of undecided voters. It would be useful to empirically evaluate the validity of our assumption, and the comparative accuracy of alternative strategies for treating undecided voters. If pollsters were seriously concerned about this assumption, they could suggest a different allocation algorithm (based upon partisanship or incumbency, for example), or adopt an allocation method of their choice when publishing pre-election estimates.

Third, the measure "standardizes" for the actual election result, providing a measure of bias that is comparable over elections for different offices in different states. The magnitude of a poll's bias is defined relative to an election outcome. This makes it possible to do a meta-analysis of the nature and causes of bias affecting an entire corpus of polls conducted for different elections in different years. Below, we illustrate this by applying the measure to several well-known presidential elections. In section 4, we use it to conduct a meta-analysis of state polls conducted for the 2002 gubernatorial and senatorial races. This measure can also be used to compare the performance of individual polling firms or polls across a number of elections or races, and it could also be applied to nonpartisan elections such as referenda.

In addition to offering several advantages, the new measure differs from the traditional measures in important ways that need to be understood in interpreting results. Several of its features are illustrated in Table 1, which compares calculations of the new measure and the traditional measure 5 for several (real and hypothetical) polls in two 2002 races.

Table 1. A Comparison of Values of *A* and Mosteller's Measure 5 for Real and Hypothetical Polls for the 2002 Alabama and New York Gubernatorial Races.

Race	Election or Poll Outcome	Odds (R/D or r/d)	(r/d)/ (R/D)	A (s.e.)	Measure 5 = |(r-d) – (R-D)|
AL Gov.	50.1% (R) 49.9% (D)	1.005			
Poll 1	45% (r) 39% (d) 16% (u)	1.15	1.15	.138 (.073)	|6-0| = 6
Poll 1*	53.6% (r) 46.4% (d)	1.15	1.15	.138 (.073)	|7.2-0| = 7.2
NY Gov.	49.9% (R) 34.5% (D) 15.6% (I)	1.445			
Poll 2	48% r 27% d 17% i 8% u	1.78	1.23	.207 (.083)	|21-15.4| = 5.6
Poll 3	53% r 39% d 2% i (or o) 6% u	1.36	.94	−.061 (.096)	|14-15.4| = 1.4

Poll 1 is the same one used for illustrative purposes so far, and poll 1* is identical to it, except the undecided have been eliminated and the percentage preferring the Republican and Democratic candidates recalculated on the new base. (This assumes, as discussed above, that undecided voters are proportionally distributed.) Note that the poll odds, and *A*, are unaffected by this exclusion, but the value of Measure 5 increases. In general, Measure 5 changes as the fraction of undecided fluctuates, even if the ratio r/d remains constant. Assuming r/d remains constant, our measure is insensitive to

fluctuations in the size of the undecided category. This is not true for Measure 5 or other measures based on the absolute value of percentage point differences. This is a positive feature of the measure, because methodological factors (e.g. the format of the preference question; see McDermott and Frankovic, 2003) can influence the number of voters who say they are undecided.[5]

Comparing the accuracy of poll 1 (for the very close Alabama governor's race) and poll 2 (for the landslide New York governor's race) illustrates another important feature of the new measure. Measure 5 suggests that poll 2 was slightly more accurate than poll 1, while A would lead to the opposite conclusion, because poll 1 has a lower value of A. The old measure measures absolute differences (and errors) while the new measure is concerned with relative differences. Another way to think about the new measure is that a given percentage point difference between two candidates in a poll is "larger" (in a relative sense) in a close election than in a blowout election.

A third point is illustrated by comparing polls 2 and 3. Measure 5 shows poll 3 to be more accurate than poll 2. Our new measure shows the same thing, but further reveals that they are "biased" in the opposite direction, with poll 2 considerably overstating Republican strength and poll 3 slightly understating it. Our measure facilitates inferences and statistical comparisons about the direction and degree of bias that are not possible using the old measures[6].

Finally, it is important to note the particular sense in which we interpret A_{ijk} as a measure of accuracy: A_{ijk} measures the accuracy of a poll *as a predictor of an election result*. A poll result might not accurately reflect voters' relative preferences between the Republican and Democratic candidates for several reasons. One reason is sampling error. Flaws in survey design might contribute to a poll's failure to predict an election outcome if, for example, a sample was not designed to represent the participating electorate on Election Day.

However, public preferences probably shift during a campaign. A poll that perfectly measured preferences at the time of the poll might not predict the eventual outcome of an election due to changes in the electorate, not a flaw of the poll. Election campaigns have a dynamic that causes voter sentiment to shift over the course of the campaign, and polls legitimately seek to measure this shifting sentiment. Polls measuring early sentiment should not be considered inaccurate if early sentiment changes as a result of the campaign. We refer to A as *predictive accuracy* to emphasize the particular sense in which we interpret it as a measure of accuracy.

3. Re-Assessing Past Presidential Polls

We first illustrate our new measure of predictive accuracy by applying it to characterize the well-studied 1948, 1996, and 2000 presidential elections. In each case, we average the mean value of A over the polls conducted for that election, treating each poll as a single (unweighted) observation to calculate the standard error of $\bar{A}$.[7]

Table 2. Mean Predictive Accuracy of Polls for Three Presidential Elections

Election	$\bar{A}$	s.e. of $\bar{A}$	N of polls
1948	.2783	.0781	3
1996	−.0838	.0221	9
2000	.0630	.0121	19

Sources of data: Mosteller et al. (1949), p. 17; Table 1 in Mitofsky (1998); Table 1 in Traugott (2001).

Table 2 shows that $\bar{A}$ for the 1948 election is significantly positive (more than three times its standard error), consistent with the familiar fact that the election polls that year showed a spectacular Republican bias. $\bar{A}$ for the 1996 presidential election is significantly negative, showing a Democratic bias, as Ladd (1996) charged and Mitofsky (1998) affirmed using a cruder measure of bias. Finally, $\bar{A}$ for the 2000 presidential election is significantly positive, showing a Republican bias. This is consistent with the fact that 14 of 19 pre-election polls analyzed indicated George Bush would win the popular vote.

We compared our new measure with measures proposed by Mosteller using Traugott's (2001) assessment of poll accuracy in the 2000 election. We ranked the accuracy of the 19 pre-election polls by three measures—predictive accuracy (A), and Mosteller's Measures 3 and 5. (For our measure of predictive accuracy, we ranked polls according to how close the absolute value of *A* is to 0.) The correlation between rankings on the two Mosteller measures is .77, while the correlations are .81 between rankings on Measure 3 and A and .97 between rankings on Measure 5 and A. It is not surprising that the second correlation is higher because Measure 5 is the absolute difference between the two leading candidates compared to their division of the vote, conceptually closer to the log odds ratio. The data used to construct the Mosteller measures include an allocation of the "undecided" portion of the sample for Measure 3. Thus, when used to rank individual polls in an election in which most were "biased" in the same direction, *A* provided consistent information with the traditional measures, especially Measure 5.

[5] The extent to which the odds r_i/d_i remains constant in the presence of fluctuations in the size of the undecided category should be empirically addressed.

[6] Alternatively, Measure 5 might be modified by taking the signed value of the differences, rather than their absolute value. Our measure is more amenable to log linear analysis, as noted below.

[7] Standard errors for $\bar{A}$ were calculated using a jackknife replication method in VPLX (Fay, 1998) and treating each election year's polls as simple random samples.

4. Did the 2002 Pre-Election Polls Support Inferences?
The number of distinct pre-election polls is quite limited in presidential elections, and each estimate is assessed against the same outcome. The advantages of evaluating the measure of predictive accuracy in non-presidential statewide elections are the greater number of polls and more varied range of outcomes. One of Mosteller's most compelling and pessimistic messages was his concern about the difficulty (approaching impossibility) of predicting close state elections (Mosteller et al., 1949:70). So we first ask whether election predictions are based on sound statistical inferences from the poll results. At the most elementary level, we ask whether a poll can support any inference at all about a leader or likely winner. To address this question, we first calculate margins of error for each poll, applying the assumption of simple random sampling. We compare them to reported margins of error, and use them to calculate confidence intervals and assess whether any projection of a winner was supportable from the poll. We then look at those cases in which a projection was supported to see if the poll predicted the eventual winner.

We recalculated the percentages identified by a poll as probable voters for the Democratic (p_D) and Republican (p_R) candidates, excluding undecided or third party voters from the denominator. The standard error is calculated as the square root of $(p_D * p_R)/n$, where n is the total number in the sample identified as probable voters for the Democratic or Republican candidates (thus we do not allocate undecided or third party voters, which seems to us to artificially inflate the sample size unless the allocation is based upon information about those individuals' preferences, which we do not have). We calculate a margin of error equal to 1.96 times the standard error, and a 95% confidence limit around the percentage Republican = $p_R \pm 1.96$ times the standard error.

The mean reported margin of error is 4.09 and the mean actual margin of error is 4.36, based on the 484 polls that reported a margin of error and provided sufficient information to calculate it. Thus, we conclude that the state polls slightly understate their margins of error calculated as described above, but the difference is not statistically significant. We probably obtain slightly higher standard errors because we are thinning the sample by dropping undecided voters from our calculations.

Its confidence interval permits us to assess whether a poll supported *any* inference—correct or incorrect—about a likely winner. Table 3 includes 504 polls for which confidence intervals could be calculated. It shows poll projections for close races (defined as a 52%-48% or closer final vote split) and for races won by a margin greater than 52%, separately for elections won by Democrats and by Republicans. If its confidence interval includes 50%, then a poll cannot predict a majority of votes for either candidate and should conclude the election was too close to call.[8] By this criterion, 57 percent (or 286 polls) could statistically project a winner or a leader, while 43 percent (218 polls) could not, as shown in the "total" row of Table 3. Of the polls that could project a winner, the projection was correct 95% of the time (54% of the polls correctly projected the winner, and 3% were in error). When they could support projections, the polls were highly accurate in all elections, except in close elections won by Republicans.

Table 3. Poll projections, by election outcome

Election Outcome	Yes—projects winner		No—too close to call	Total
	Correctly	In error		
Dem. win by 53% or more (N=148)	74%	1	26	100%
Dem. win in close race (N=49)	18%	2	80	100%
Rep. win in close race (N=41)	7%	12	80	100%
Rep. win by 53% or more (N=266)	57%	3	41	100%
Total polls (N=504)	54%	3	43	100%

As would be expected, polls in tight races were less able to sustain statistical projections than polls in races with a greater margin of victory for either party: 80% were "too close to call" statistically. Even when the margin of victory was greater than 52-48, many polls were too close to call (26% in elections won by Democrats, and 41% in races won by Republicans). The large fraction of polls that could not statistically support a projection in such elections suggests that sample sizes are too small for this purpose.

5. Were the 2002 State Pre-Election Polls Unbiased?
In the absence of overall bias, the mean value of the A_{ijk} averaged over all 548 state polls would be expected to be 0. However, the mean value of A_{ijk} is –.0330, with standard error .0077. This implies a small but statistically significant Democratic bias over the polls as a whole. Although quite small, the bias is potentially important in close races: if all races were perfectly tied, on average the polls would have estimated a Republican share of 49.18% rather than 50%.
We may analyze the new measure to assess potential sources of bias, including partisan poll auspices and methodological factors. In the meta-analysis conducted below, we treat each

[8] We do not examine whether a poll release was accompanied by a statement that a race was "too close to call," since we do not have the reports that accompanied the release of these polls. It would be worthwhile to examine whether published discussions of the poll results were consistent with estimates of statistical uncertainty.

poll as a single (unweighted) observation.[9] We might obtain different results if we weighted by sample size, although the differences would probably not be too great since the variability in the number of interviews is not too large (ranging from about 300 to 1500 among 548 polls, with a mean of 484, excluding the undecided).

a. Bias of the auspices. We examine the accuracy of state polls according to their partisan auspices in Table 4.

Table 4. Mean predictive accuracy Ā, by partisan auspices (standard errors in parentheses)

	Nonpartisan	Democratic poll[1]	Republican poll
Ā	−.0304	−.1576	.0699
	(.0082)	(.0241)	(.0270)
N of polls	469	41	38

[1]Includes one poll with dual auspices. Standard errors are calculated with jackknife replication methods in VPLX (Fay, 1998).

Accuracy is significantly different from zero for polls in all three categories, as reflected by estimates of A more than twice the standard errors in each case. The direction of bias is different, however.

Polls with no partisan affiliation were slightly biased in the Democratic direction, as indicated by a significant negative value of Ā. (Ā = −.0304 implies that these 469 polls would have estimated 49.24% Republican vote, on average, in perfectly tied races.)

Partisan polls are extremely biased. Democratic polls are biased in favor of Democratic candidates, and Republican polls are biased in favor of Republican candidates, relative to election results. The difference between the two values of A is highly significant (t=6.3), and both Democratic and Republican polls are also significantly more biased than nonpartisan polls.

The value of Ā = −.1576 for Democratic polls would translate into an average estimate of 45.97% preferring Republican candidates in perfectly tied races. Ā = .0699 for Republican polls implies an average estimate of 51.75% preferring Republican candidates in perfectly tied races.

Partisan biases might have many sources. Democratic and Republican pollsters may use different methods that favor their party's candidates. Partisan pollsters might release polls selectively, so that Democratic pollsters only released their polls publicly if they are favorable to Democratic candidates, while Republican pollsters only released polls favorable to Republicans. In addition, the partisan and nonpartisan polls occurred in different types of races. Partisan polls were concentrated in certain states (Louisiana, New Hampshire, South Dakota, Tennessee, and Texas), were more frequent in gubernatorial races than Senate races, and tended to be early rather than late in the campaign. Some differences in predictive accuracy shown in Table 4 may reflect differences in the particular campaigns in which partisan polling occurred. Because the partisanship of a poll appears to be such a potentially large source of bias, we control for it in the tables below.

b. Methodological influences. The methods used in a poll may influence its predictive accuracy. It is commonly believed that polls taken close to an election are more accurate than those taken far in advance. (Thus, the NCPP included only final polls in its review, and dropped polls in which interviewing was completed before October 20th.) Different polls frame preference questions differently (see e.g., McDermott and Frankovic, 2003) and may rely on different methods for identifying people who will actually vote in an election. Different polls might use different sample designs that might affect results. Here, we consider just two possible methodological influences on poll accuracy: the number of weeks in advance of the election a poll was taken, and a poll's reliance on likely voters or registered voters to project a winner.

Table 5 shows the mean predictive accuracy by the number of weeks in advance of the election a poll was taken. Results are shown separately for nonpartisan, Democratic, and Republican polls.

Table 5. Mean predictive accuracy Ā of state polls, by number of weeks poll was taken in advance of an election

Poll auspices	5 –10 weeks before	4th week	3rd week	2nd week	During final week
Non-partisan polls	.0115	−.0332	−.0889	−.0350	−.0453
	(.0176)	(.0361)	(.0266)	(.0192)	(.0104)
Dem.	−.1321	−.2373	−.2087	−.1259	−.0873
	(.0392)	(.0805)	(.0323)	(.0454)	(.0267)
Rep.	.0762	.0731	.0755	.0360	.0625
	(.0331)	(.1096)	(.0820)	(.0617)	(.0288)

In general, polls did not predict the final vote better the closer they were to the 2002 election. Nonpartisan polls taken more than a month in advance were significantly more accurate (Ā is closer to zero) than those taken in the final week (t = 2.78). A significant Democratic bias emerged in neutral polls the last three weeks of the campaign, as shown by negative values of A that are twice their standard errors. This result contrasts with Crespi's (1988) finding of a slight positive correlation between accuracy and timeliness. It may be consistent with Gelman and King's (1993) argument that the ups and downs of candidate popularity during the course

[9] It is important to note that there may be dependence among different polls that may affect the results of our analysis. For example, different polling firms may use different sample designs that systematically affect their results. Selective release (as described below) would also create dependence.

of a campaign may be irrelevant to the final outcome, which may have more to do with stable attitudes and partisan preferences that preceded the campaign.

Democratic polls show a significant Democratic bias throughout the period. The magnitude of the bias appears to decline over time; A is significantly larger the fourth week out than during the final week.

Republican polls show a consistent Republican bias throughout, although it is significantly different from zero only in polls 5 to 10 weeks in advance of the election, and in the final week.

These results cast some light on why the final election results were surprising to many poll watchers. *Nonpartisan polls significantly overstated Democratic strength during the last month of the campaign*. While the reasons for this pattern are beyond the scope of this paper, they merit further analysis of methodological factors as well as election dynamics.[10] One possible methodological source of bias is examined in Table 6, which shows the predictive accuracy of polls of likely voters and registered voters. (It also includes polls for which no information was available about the method of identifying voters.)

Table 6. Mean predictive accuracy Ā of state polls, by voter identification method and partisan auspices of poll

Poll auspices	Likely voter	Registered voter	General population	Unknown
Nonpartisan	−.0409 (.0100)	.0285 (.0233)	−.1692[1]	−.0330 (.0182)
Democratic	−.1626 (.0251)	−.0512[1]	___	−.0684[1]
Republican	.0568 (.0293)	.1813 (.0304)	___	___
N of polls	393	68	1	86

[1]N = 1, so no standard error can be calculated.

Likely voter polls are more Democratic than registered voter polls. One poll of the general population produced an extreme Democratic bias. These methodological differences are well known: the general population is more Democratic than the electorate because Republicans are more likely to vote than Democrats. The difference between likely voter and registered voter polls holds regardless of the partisanship of the pollster. However, differences in their methods does not account for the extreme biases of partisan polls. Both Republican and Democratic polls are biased in favor of the party's candidates, regardless of whether likely voters or registered voters were polled.

Likely voter polls by nonpartisan pollsters show a significant Democratic bias, with registered voter polls unbiased (i.e., Ā is not significantly different from 0). It is likely that identification of registered voters (who are more likely to be Republican) overcame the slight Democratic bias that apparently characterized nonpartisan polls in the 2002 state elections.

The most extreme biases occurred in likely voter polls by Democratic pollsters, registered voter polls by Republican pollsters, and a general population poll.

Conclusions

Our new measure of predictive accuracy appears to be useful for characterizing the accuracy of the pre-election polls as well as the extent and direction of any biases they produce. It has the considerable advantage over the Mosteller measures of providing a summary statistic that is comparable in different elections and for early as well as late polls, thus lending itself to meta-analyses of the sort conducted here. It could also be used to compare the performance of polling organizations across elections and across races, although we favor more focus on understanding the predictors of election forecasting rather than differences in results between different organizations. The measure of predictive accuracy could also be applied to study referenda elections.

Our meta-analysis confirms a small but statistically significant overstatement of Democratic strength in nonpartisan polls conducted in state elections in 2002. The polls conducted for the presidential election of 1996 also slightly overstated the Democratic advantage compared to final election results. This result contrasts with the pre-election polls in the 2000 presidential race, which significantly overstated the Republican lead compared to final election results. These differences in poll performance in different elections point to campaign dynamics and election-specific factors as potential influences on the predictive accuracy of polls, and suggest that important influences on election outcomes, and the accuracy of the polls, remain unidentified.

Our meta-analysis provides conclusive evidence that partisan polls lack credibility. Both Democratic and Republican polls produced extremely biased results that favored their party's candidates. We do not know whether the biases arise from selective publication of favorable results, or from actual differences in the methods used; both possibilities merit further investigation. Based on our results, we can recommend that readers should ignore such polls, and journalists should take their evident biases into account when declining to report them.

Identifying the possible causes of bias is beyond the scope of this paper, but we note that the methods used to conduct

[10] While one possible explanation of bias in the final pre-election polls might have been their inability to pick up the Republican mobilization efforts through the 72-Hour Task Forces, this cannot be the complete explanation because the biases appeared early in the campaign, before their work had started.

the polls may be contributing factors. In 2002, likely voter polls were more Democratic and registered voter polls were more Republican, while general population polls were the most Democratic of all. However, the data do not suggest that differences in methods account for the extreme biases of partisan polls, which persist after controlling for whether polls targeted likely voters or registered voters. Analysis of the influence of other methodological differences might be fruitful. Investigation of the methodological and statistical underpinnings of election forecasts seems warranted.

Our analysis does not confirm the common belief that timely polls more accurately predict election outcomes than early polls. Indeed, our data show that polls conducted by nonpartisan pollsters more than a month before the 2002 election were less biased predictors of election outcomes than polls conducted the week before. This finding may reflect the dynamics of the 2002 elections. In addition, past analyses of the effects of poll timeliness have relied on a measure of error that did not take into account the direction of errors, as ours does, and that may account for the difference in findings.

We believe that our new measure can prove useful as a summary measure of accuracy in election forecasts. It is easily computed, summarized, and can be analyzed using multivariate statistical methods. It permits comparisons among elections with different outcomes, and among polls that vary in their treatment or numbers of undecided voters. It does not require allocations of undecided voters, although evaluation of the effects of the assumption that undecided voters split proportionally is warranted. The measure does not tell one everything one wants to know about the accuracy or bias of election forecasts. For example, it does not capture the crucial bottom line question of whether a poll "got it right" in terms of correctly forecasting the winner. We believe it can and should be used in combination with other measures of error to characterize the nature and extent of biases affecting election forecasts, and identify their sources.

References

Crespi, I. (1988). Pre-Election Polling: Sources of Accuracy and Error. New York: Russell Sage.

Fay, R. E. (1998) VPLX Program Documentation, Vol. 1. Washington DC: Census Bureau.

Gelman, A. and King, G. (1993) "Why Are American Presidential Election Campaign Polls so Variable When Votes Are So Predictable?" British Journal of Political Science 23:409-51.

Huffington, A. (1998). "Investigating the Pollsters." http://www.ariannaonline.com/columns/files/101298.html. Filed on October 12, 1998; viewed on May 7, 2003.

Huffington, A. (2001). "Some Things Never Change: The Unbearable Ludicrousness of Polling." http://www.ariannaonline.com/columns/files/101801.html. Filed on October 18, 2001, viewed on May 7, 2003.

Huffington, A. (1996). "A Modest Proposal." http://www.ariannaonline.com/columns/files/100396.html. Filed on October 3, 1999, viewed on May 7, 2003.

Ladd, E. C. (1996) "The Election Polls: An American Waterloo." Chronicle of Higher Education November 22, p. A. 52.

McDermott, M. L. and Frankovic, K. (2003) "Review: Survey Methods and Horserace Polling." Public Opinion Quarterly 67:244-264.

Mitofsky, W. (1998) "The Polls–Review: Was 1996 a Worse Year for Polls than 1948?" Public Opinion Quarterly 62:230-249.

Mosteller, F., Hyman, H., McCarthy, P., Marks, E., 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.

National Council on Public Polls Polling Review Board. (2002) Analysis of the 2002 Election Polls. Press Release issued Dec. 19, 2002, by the National Council on Public Polls, URL [http://www.ncpp.org/presspost.htm] and http://www.ncpp.org/2002SenGovPoll/2002ElectionPolls.html, visited on May 7, 2003.

Presser, S., and Schuman, H. (1980) "The Measurement of a Middle Position in Attitude Surveys." Public Opinion Quarterly 44:70-85.

Rademacher, E. W., and Smith, A. E. (2001) "Poll Call." The Public Perspective, vol. 12, no. 2 (March/April): 36.

Schuman, H., and Presser, S. (1979) "The assessment of 'no opinion' in attitude surveys." In K. F. Schuessler (ed.) Sociological Methodology 1979. San Francisco: Jossey Bass.

Traugott, M. W. (2001) "Assessing Poll Performance in the 2000 Campaign." Public Opinion Quarterly 65: 389-419.

Traugott, M. W. (Forthcoming). "Can We Trust the Polls?" Brookings Review (Summer 2003).

Turner C. F., and Martin, E. (Eds.) (1984) Surveying Subjective Phenomena. Vol 1. New York: Sage.

Visser, P. S., Krosnick, J. A., Marquette, J., and Curtin, M. (1996) "Mail Surveys for Election Forecasting? An Evaluation of the Columbus Dispatch Poll." Public Opinion Quarterly 60:181-227.