Sharon L. Lohr* and J. Michael Brick

Roosevelt Predicted to Win: Revisiting the 1936 *Literary Digest* Poll

DOI 10.1515/spp-2016-0006

Abstract: The *Literary Digest* poll of 1936, which incorrectly predicted that Landon would defeat Roosevelt in the 1936 US presidential election, has long been held up as an example of how not to sample. The sampling frame was constructed from telephone directories and automobile registration lists, and the survey had a 24% response rate. But if information collected by the poll about votes cast in 1932 had been used to weight the results, the poll would have predicted a majority of electoral votes for Roosevelt in 1936, and thus would have correctly predicted the winner of the election. We explore alternative weighting methods for the 1936 poll and the models that support them. While weighting would have resulted in Roosevelt being projected as the winner, the bias in the estimates is still very large. We discuss implications of these results for today's low-response-rate surveys and how the accuracy of the modeling might be reflected better than current practice.

1 Introduction

The *Literary Digest* (LD) poll of 1936 is a byword for bad survey research. Textbooks have long used it as a prime example of how sampling goes bad (see, for example, Parten 1950: p. 25; Hansen et al. 1953: pp. 6–7; Lohr 2010: pp. 8–9). Little (2016) included the 1936 *Literary Digest* poll in her list of the four worst political predictions in history.

The story of the 1936 poll is well known. Ten million ballots were sent out: every day more than a quarter million envelopes were addressed by hand. The mailing list was "drawn from every telephone book in the United States, from the rosters of clubs and associations, from city directories, lists of registered voters, classified mail-order and occupational data" (Literary Digest 1936a: p. 3). Of these, almost 2.4 million ballots were returned. The final results from the poll predicted that Republican Alfred Landon would win with 54 percent of the popular vote and

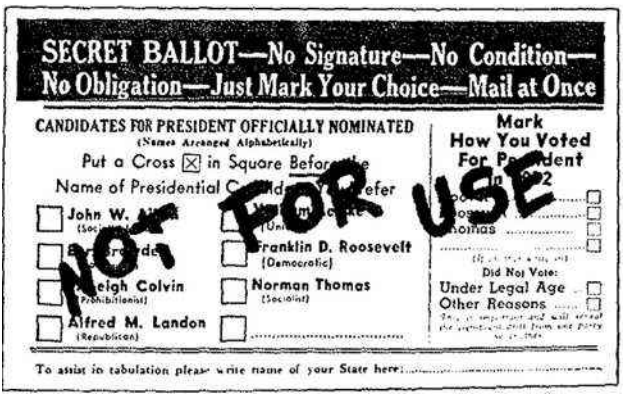
***Corresponding author: Sharon L. Lohr,** Westat, 1600 Research Boulevard, Rockville, MD 20850, USA, e-mail: SharonLohr@westat.com
J. Michael Brick: Westat, 1600 Research Boulevard, Rockville, MD 20850, USA

370 electoral votes (Literary Digest 1936c). In the election, Franklin Roosevelt won more than 60 percent of the popular vote and 523 electoral votes, carrying every state except Maine and Vermont.

Gallup (1938) blamed the poll's inaccuracy on the sampling frame, which was constructed largely from lists of telephone and automobile owners and thus overrepresented the well-to-do. Bryson (1976), however, argued that the sampling frame deficiencies did not explain the errors and ascribed the poll's failure to nonresponse bias. Other postmortems of the LD poll have also pointed to nonresponse as a primary factor in the poll's inaccuracy, relying on a survey taken by Gallup in 1937 that asked respondents whether they had participated in the 1936 LD poll. These analyses concluded that telephone and automobile owners both supported Roosevelt (though not to the same extent as persons without telephones and automobiles), and that the low response rate combined with the flawed sample produced the inaccurate forecast (Squire 1988; Cahalan 1989). We refer the reader to Lusinchi's (2012) thorough review of explanations that have been proposed in the literature for the failure of the 1936 LD poll.

But the 24 percent response rate of the 1936 LD poll is much higher than the response rate in many of today's polls. One difference is that today's polls weight the data to attempt to compensate for nonresponse bias. Typically, the weights of the respondents are adjusted so that weighted estimates match the voting population demographics; some polls also weight by political party.

Demographic weighting could not have been used for the 1936 LD poll because those data were not collected from the respondents. But the ballot (Figure 1) did ask respondents for whom they voted in the 1932 election. This information could


SECRET BALLOT—No Signature—No Condition—
No Obligation—Just Mark Your Choice—Mail at Once

CANDIDATES FOR PRESIDENT OFFICIALLY NOMINATED
(Names Arranged Alphabetically)

Put a Cross ☒ in Square Before the Name of Presidential Candidate You Prefer

John W. Aiken (Socialist Labor)
Franklin D. Roosevelt (Democratic)
Norman Thomas (Socialist)
Alfred M. Landon (Republican)

Mark How You Voted For President in 1932

Thomas
Did Not Vote:
Under Legal Age
Other Reasons

NOT FOR USE

To assist in tabulation please write name of your State here:

Figure 1: Ballot from the 1936 *Literary Digest* poll.
Source: Literary Digest (1936a: p. 3).

have been used to weight the data using techniques that were known at that time and involved simple computations.

In this paper, we explore the results of using different weighting models for adjusting the results of the LD poll, and find that poststratifying to the 1932 election results changes the predicted winner of the election. We argue that the uncertainty about the weighting model should be included in the margin of error, and draw lessons from the 1936 LD poll for today's low-response-rate surveys.

2 Weighting the *Literary Digest* Poll Results: What Was Known in 1936

In 1936, statisticians knew about weighting as a technique for adjusting survey results. Laplace (1814) had used ratio adjustments to estimate the population of France in 1802. Bowley (1926) had used weighting methods with purposive samples and Neyman (1934), discussing the purposive sampling of Gini and Galvani, had used weighted means to describe their estimation procedure.

Weighting and other adjustments had also been discussed for the LD polls. Robinson (1932) adjusted the results of the 1928 LD poll by the error in the 1924 poll. He used an additive adjustment which he called correcting the plurality error. He defined the plurality error for a state s in 1924 as PE (1924, s) = (LD 1924 percentage for Republican candidate – LD 1924 percentage for Democratic candidate) – (actual 1924 percentage for Republican candidate – actual 1924 percentage for Democratic candidate) in state s. Robinson then subtracted PE (1924, s) from the LD 1928 estimate of the plurality in each state s. For the 1928 election, Robinson concluded that this method reduced the average state plurality error by 6 percentage points.

Crum (1933) argued that the best type of sample to use for a straw poll would be a random sample, and that polls should be conducted to allow assessment of their representativeness with respect to political party, geographical distribution, sex, race, and economic status. He used the information in the LD 1932 poll on how respondents had voted in the 1928 election to project the 1932 results. Crum calculated that 57.7 percent of those who had reported voting Republican in 1928 said they were planning to vote for Hoover in 1932; the corresponding Hoover percentages were 9.2 percent for those who had reported voting Democratic in 1928 and 39.2 percent for those who had reported not voting in 1928. Then, assuming that 90 percent of the 1928 voters would vote in 1932, and that the "no vote" group from 1928 would comprise 15 percent of the 1932 electorate, Crum projected

that Hoover would get 33 percent of the popular vote in 1932. He did this same projection for every state to predict the electoral college result.

The same day – October 31, 1936 – that the *Literary Digest* published its final projections for the election, the *New York Times* published a letter to the editor signed by "Amateur Statistician." Amateur Statistician (1936) noted that less than one-fourth of the LD 1936 ballots were returned and wrote: "This fourth is, so to speak, self-chosen. It represents only about 5 per cent of the total probable vote. In so small a sampling the possible margin of error is high. If, moreover, as many contend, large city populations, industrial labor or the lower income groups were greatly under-represented in the ballots that The Digest mailed out, there is no way to determine the extent of this error on the face of the returns." The letter then mentioned some of the other weighting adjustments that had been proposed, including Crum's (1933) method and an adjustment based on changes in sentiment in 1936 suggested by Franklin (1936), and again described the possibility of using the reported vote in the previous election. Levy (1936) also stated that the poll results should be corrected by the shift among voters who had reported voting for Hoover in the 1932 election, and concluded that such a correction would affect the outcome in eight states.

3 Weighting Adjustments for the *Literary Digest* Poll of 1936

All the methods that had been suggested for adjusting the 1936 LD poll correspond to different models for the poll's coverage and response mechanisms. Even using the unweighted results corresponds to a model that noncovered segments of the population and nonrespondents vote similarly to the respondents to the poll. Robinson's (1932) plurality error method assumes that the bias towards the Republican candidate is additive and is the same for a state in successive elections.

To explore the effect of weighting by the vote reported on the prior election similar to the Crum (1933) suggestion, we use a simple ratio adjustment for the weights of respondents to the poll, separately for each state. Cornfield (1942) used a similar ratio adjustment for the national vote in his famous article on nonresponse bias in surveys. The table of state-by-state results in Literary Digest (1936c) gave the number of respondents supporting each of the three major candidates (Landon, Roosevelt, and Lemke) in the 1936 poll, and also recorded how the same respondent reported voting in the 1932 election. The actual election results from 1932 are used as the control totals for weighting the respondents to the LD 1936 poll.

Let x_{cs} denote the number of poll respondents supporting candidate c (Landon, Roosevelt, or Lemke) in state s in 1936, and let x_{csj} denote the number of those respondents who report voting in category j in 1932. The categories used for j are Republican (R), Democratic (D), Socialist or Other (S), Did not vote in 1932 (N), and Missing 1932 vote (M).

If the poll were representative and if respondents accurately reported their 1932 voting behavior, we would expect the percentage of respondents listing that they voted as R, D, or S in 1932 to be close to the actual 1932 election results. We therefore define the weighting adjustment for category j (for $j = R$, D, or S) in state s to be w_{sj} = (percentage vote for category j in state s in 1932 election results)/(percentage of respondents from state s who reported category j for 1932, among those who reported voting in 1932).

We have no information on the persons in categories N and M for the 1932 election. We assume, for lack of better information, that they are distributed in proportion to the respondents who did report their 1932 vote. Thus, we define $w_{csN} = w_{csM} = (w_{sR}x_{csR} + w_{sD}x_{csD})/(x_{csR} + x_{csD})$. An alternative definition could include the Socialist or Other category in this apportionment, but the results are essentially the same with the alternative definition.

Table 1 displays the weighting adjustments used for the 1936 LD poll, sorted by the weights for the persons who had reported voting Democratic in 1932. The weighting adjustments for the states indicate that persons who said they had voted Republican in 1932 were overrepresented in the 1936 data for all but five states (all in the South), and persons who said they had voted Democratic or for another candidate in 1932 were underrepresented in the 1936 data for all but five states.

With these weight adjustments, the estimated relative count for candidate c in state s is

$$y_{cs} = \sum_{j=R,D,S} w_{sj}\, x_{csj} + \sum_{j=N,M} w_{csj}\, x_{csj}. \tag{1}$$

Using equation (1), the estimated percentage for Roosevelt in state s is

$$100 \times y_{\text{Roosevelt},s} / (y_{\text{Roosevelt},s} + y_{\text{Landon},s}).$$

The actual percentage for Roosevelt in 1936 is similarly calculated as 100× (vote for Roosevelt)/(total for Roosevelt and Landon).

The model for the ratio-weighted prediction used in equation (1) assumes that persons reported their 1932 vote accurately, and that the persons who were not in the survey (either through undercoverage or nonresponse) had the same relationship between the 1936 vote and 1932 vote as the persons who responded

Table 1: Relative weights for 1936 *Literary Digest* poll respondents based on 1932 reported vote.

State	State Abbr.	Relative Weight for 1936 LD Respondent with 1932 Vote for				
		Republican	Democrat	Other	Missing/no vote (Landon 36)	Missing/no vote (Roosevelt 36)
Massachusetts	MA	0.66	1.77	2.36	0.80	1.52
Rhode Island	RI	0.65	1.72	1.52	0.79	1.51
New Hampshire	NH	0.72	1.70	0.83	0.84	1.51
Vermont	VT	0.82	1.40	1.88	0.89	1.27
Connecticut	CT	0.75	1.40	2.75	0.83	1.26
New Jersey	NJ	0.76	1.40	1.70	0.86	1.25
Michigan	MI	0.74	1.37	1.94	0.86	1.23
South Dakota	SD	0.66	1.37	1.85	0.85	1.22
Delaware	DE	0.81	1.31	1.34	0.87	1.17
Colorado	CO	0.74	1.30	1.96	0.84	1.20
Kansas	KS	0.76	1.30	2.45	0.87	1.18
Iowa	IA	0.73	1.30	2.14	0.85	1.19
Montana	MT	0.68	1.30	3.00	0.80	1.17
Nebraska	NE	0.70	1.29	1.54	0.85	1.20
Wisconsin	WI	0.64	1.27	3.30	0.81	1.17
Minnesota	MN	0.71	1.27	2.57	0.83	1.16
North Dakota	ND	0.64	1.27	2.35	0.82	1.14
California	CA	0.71	1.27	2.78	0.82	1.14
Wyoming	WY	0.75	1.27	2.60	0.86	1.18
Nevada	NV	0.68	1.27	0.00	0.85	1.16
Idaho	ID	0.73	1.26	2.49	0.84	1.17
Indiana	IN	0.78	1.25	2.74	0.87	1.16
Oregon	OR	0.73	1.23	2.49	0.82	1.13
Illinois	IL	0.78	1.23	2.11	0.89	1.13
Missouri	MO	0.76	1.21	1.22	0.87	1.15
Arizona	AZ	0.72	1.19	1.65	0.87	1.13
Washington	WA	0.67	1.19	6.13	0.80	1.10
Oklahoma	OK	0.73	1.16	0.00	0.88	1.12
Maine	ME	0.90	1.15	1.37	0.94	1.12
Ohio	OH	0.86	1.15	1.86	0.91	1.09
Louisiana	LA	0.41	1.13	0.28	0.81	1.08
West Virginia	WV	0.88	1.12	2.10	0.93	1.09
New Mexico	NM	0.88	1.08	1.32	0.94	1.05
Texas	TX	0.65	1.07	1.31	0.87	1.05
New York	NY	0.87	1.07	2.41	0.91	1.04
Florida	FL	0.85	1.06	0.54	0.94	1.05
Georgia	GA	0.67	1.04	1.29	0.89	1.03
Arkansas	AR	0.76	1.04	2.24	0.88	1.03
Pennsylvania	PA	0.92	1.04	3.17	0.95	1.01
Utah	UT	0.93	1.03	2.02	0.96	1.01
South Carolina	SC	0.41	1.03	0.37	0.88	1.02

Table 1 (continued)

State	State Abbr.	Relative Weight for 1936 LD Respondent with 1932 Vote for				
		Republican	Democrat	Other	Missing/no vote (Landon 36)	Missing/no vote (Roosevelt 36)
Mississippi	MS	0.61	1.02	2.65	0.85	1.01
Maryland	MD	0.95	1.02	1.61	0.97	1.01
Alabama	AL	1.02	0.99	2.14	1.01	0.99
Virginia	VA	1.03	0.98	1.44	1.01	0.99
Kentucky	KY	1.05	0.96	1.27	1.03	0.97
Tennessee	TN	1.14	0.94	1.17	1.08	0.96
North Carolina	NC	1.34	0.90	0.95	1.20	0.93

to the survey. Wright (1993) and Campbell (2010) found evidence that survey respondents who are matched with voting records may report having voted for the winning candidate in the election when they actually voted for a different candidate, but the effect of such misreporting for presidential elections was small – about 1–1.5 percentage points for most presidential elections. The ratio-weighted model also assumes that 1936 respondents were accurate in their reports that they voted in 1932; although some analysts have concluded that discrepancies between post-election surveys and government voting records are evidence that respondents have misrepresented their voting history (Belli et al. 2001), Presser et al. (1990) and Berent et al. (2016) argued that some of those discrepancies are due to record linkage errors.

Figure 2A shows the actual and unweighted predicted percentages for Roosevelt, for each state. Figure 2B displays the predicted percentages calculated using equation (1). The weighted and unweighted estimates both underestimate the percentage of the vote actually captured by Roosevelt. But the weighted estimates are closer to the actual results, and most states have a higher percentage for Roosevelt with the weights than without the weights. Crucially, with the weights, 10 states move from Landon to Roosevelt, including the high-electoral-vote states of California and New York. The unadjusted *Literary Digest* results predicted that Roosevelt would win 16 states and 161 out of 531 electoral votes. With the ratio-weighting adjustments from equation (1), Roosevelt is predicted to win 26 states and 276 electoral votes, capturing 52% of the electoral votes. Thus, this weighting adjustment predicts the correct winner of the election, although it still does a poor job of predicting the margin of the win in terms of the popular vote and the electoral college.

Figure 2C and D present two other weightings that were possible using known methods and the available data. Figure 2C applies the additive error correction

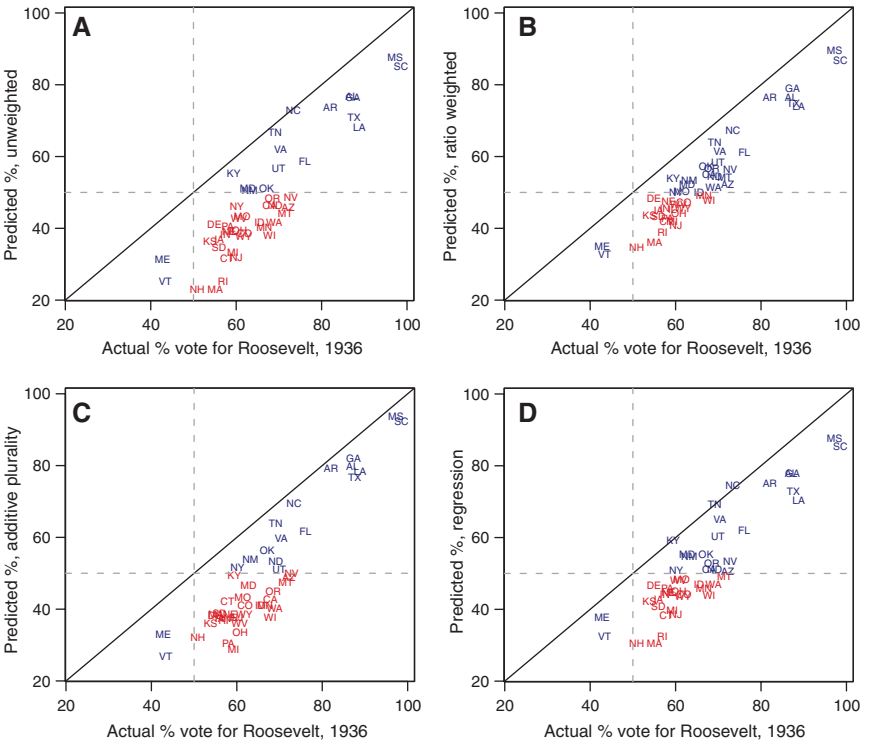


Figure 2: Predicted percentage of vote for Roosevelt in the 1936 election, using (A) unweighted counts from *Literary Digest* poll, (B) ratio adjustment, (C) Robinson's (1932) additive plurality adjustment, and (D) regression prediction.
Note: The states in blue (upper right and lower left quadrants) are those for which the poll predicted the correct winner of the state. The states in red (lower right quadrant) are those for which the poll predicted the wrong candidate would win.

proposed by Robinson (1932), where the error in the predicted percentage for Roosevelt in 1932 is added to the unweighted predicted percentage for Roosevelt in 1936, separately for each state.

Figure 2D displays the predicted vote from a regression prediction. The regression model

$$\text{Actual vote for Democrat} = 10.2155 + 0.8831\,(\text{unweighted predicted vote for Democrat})$$

($R^2 = 0.89$) was determined using the 96 state observations in the 1932 and 1928 elections, and the predicted vote for Roosevelt in 1936 was determined by finding

the predicted values from this regression equation using the unweighted 1936 vote as the independent variable. The regression prediction assumes that the understatement of Democratic votes is the same for 1936 as for the previous two elections. We also performed a regression adjustment based on only the 48 observations in the 1932 election; that model has intercept 3.66 and slope 0.97 and the 1936 predictions were very similar to the unweighted results for 1936.

4 Why Didn't the *Literary Digest* Weight the Data?

The Literary Digest (1936c: p. 5) editorial board was aware of weighting techniques but did not use them: "These figures are exactly as received from more than one in every five voters in our country – they are neither weighted, adjusted nor interpreted." The *Digest* viewed its role as an impartial presenter of data. For the 1928 poll, Literary Digest (1928a: p. 11) wrote: "Here, as always, the DIGEST itself is acting as a mere recorder of opinion…. It presents its figures, vouches for the honesty and carefulness with which the poll was taken, and leaves its readers to draw their own conclusions."

After the election results showed the 1936 LD poll to be horribly wrong, the editorial board again defended its choice not to weight: "Figures – So the statisticians did our worrying for us on that score, applying what they called the "compensating-ratio" in some cases, and the "switch-factor" in others. Either way, for some of the figure experts, it didn't matter; interpret our figures for 2,376,523 voters as they would, the answer was still Landon. Then other statisticians took our figures and so weighted, compensated, balanced, adjusted and interpreted them that they showed Roosevelt. We did not attempt to interpret the figures, because we had no stake in the result other than to wish to preserve our well-earned reputation for scrupulous bookkeeping" (Literary Digest 1936d: pp. 7–8). They were also aware that their polls had "*always* got too big a sampling of Republican voters" from 1920 onward (Literary Digest 1936d: p. 7) yet the poll results had predicted the correct winner in the past.

A large part of the reason that the *Literary Digest* chose not to weight the results was that the unweighted results had worked well for the 1932 and 1928 elections. They also noted that there was historical precedent for expecting Roosevelt to have fewer votes in 1936 than he had in 1932 because in previous elections the vote for a candidate was lighter for a re-election than for a first election (Literary Digest 1936d). They thus concluded that their model from previous elections – namely, that the respondents to the poll were representative of voters in the 1936 election – was likely to be valid in 1936 as well.

The accuracy of their assumption for the 1932 election can be seen in Figure 3A, which shows the actual percentage voting for Roosevelt and the unweighted predicted percentage from the poll in 1932, using data from Literary Digest (1932b). Figure 3B shows the result of using the weighting in equation (1) for the 1932 election, adjusting results by the respondents' self-reported vote in the 1928 election. Overall, the weighting does not appreciably improve the estimates for the outcome of the 1932 election. The weighted percentages for some of the southern states such as Arkansas and Alabama are further from the actual

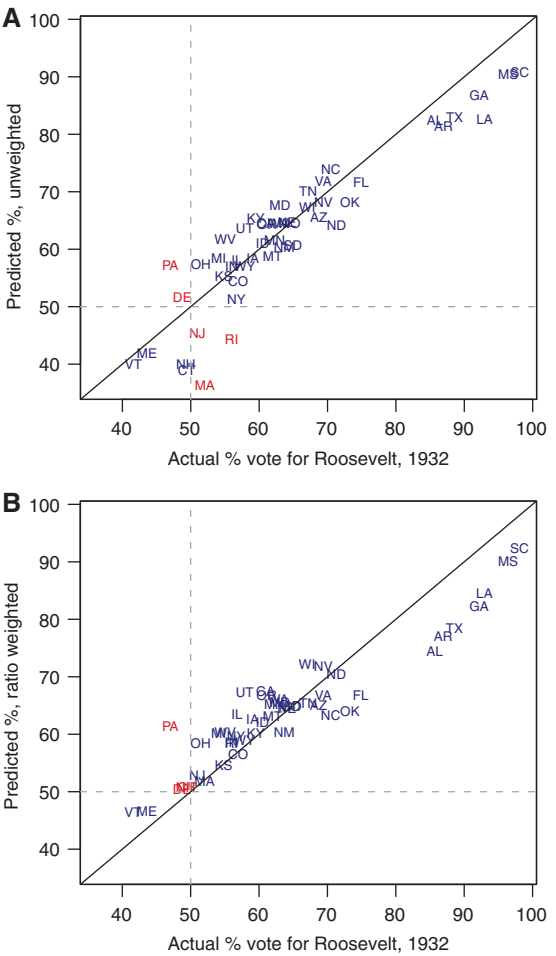


Figure 3: Predicted percentage of vote for Roosevelt in the 1932 election, using (A) raw returns from *Literary Digest* poll, and (B) weighted estimates from *Literary Digest* poll.

results than the unweighted percentages, but for most states the weighting does no harm. The weighting improves the estimates for several of the states in the Northeast.

Figure 4 shows similar results for the 1928 election (Literary Digest 1928b), with the unweighted percentages in Figure 4A and the weighted percentages, adjusting for the reported votes in the 1924 election, in Figure 4B. Even though the 1924 election was unusual in having a third-party candidate, Robert LaFollette, who received approximately 17 percent of the popular vote in the election (and it

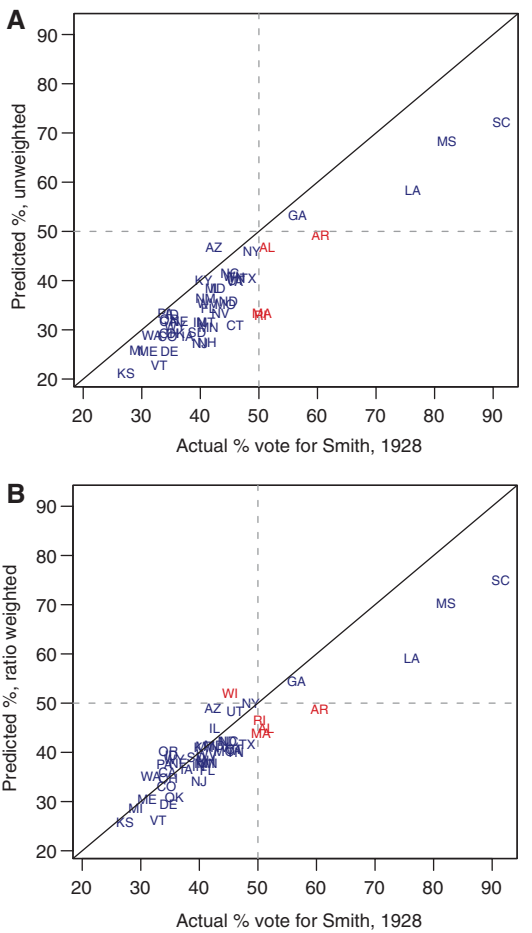


Figure 4: Predicted percentage of vote for Smith in the 1928 election, using (A) raw returns from *Literary Digest* poll, and (B) weighted estimates from *Literary Digest* poll.

is thought that many of those votes were from registered Democrats), the weighting method improves the prediction of the 1928 election results.

Figures 2–4 all have an underrepresentation of votes for the Democratic candidate in southern states such as South Carolina and Mississippi. In 1928, the weighting adjustments corrected for the bias in most of the remaining states, but not in the states with the highest percentages for Smith, which were all in the South. For the southern states, the weighted estimates were further away from the election results than the unweighted estimates. Walton et al. (2001) suggested that the inaccuracy of the LD poll in the South may have been due in part to African Americans who were represented in the sampling frame but not among the voters.

5 Accounting for Uncertainty in the Weighting Adjustments

The results from the 1932 and 1928 LD polls changed little when different weighting models were used. The results from the 1936 poll, however, changed substantially with the ratio adjustment model in equation (1). There were strong opinions in 1936 about whether weighting should be used for the poll. The *Literary Digest* editorial board strongly opposed weighting, while its rival publication *The New York Times* published many articles and letters to the editor advocating weighting.

The sampling margin of error does not capture uncertainty about the weighting model. With large sample sizes such as those for the 1936 LD poll, the sampling error is negligible for many states. The sampling margins of error for the unweighted estimates in 1936 range from 0.18 percentage points for New York to 2.3 percentage points for Nevada, with a median margin of error of 0.6 percentage points. Almost all of the uncertainty about the estimates comes from the nonsampling errors. The same problem arises when today's poll aggregators combine data and estimates and the resulting sampling errors understate the accuracy of the estimates. Lohr and Raghunathan (2017) argued that measures of uncertainty from combined data sets should include the between-source differences.

One proposed method to deal with uncertainty about the weighting model is Bayesian model averaging (Hoeting et al. 1999), which includes subjective uncertainty about the weighting models. Typically, these methods will not include uncertainty from nonsampling errors, unless the weighting models considered take these into account. Hjort and Claeskens (2003) gave a frequentist view of model averaging.

The idea behind Bayesian model averaging is simple. Let θ represent a quantity of interest (in this case, the percentage of the vote for Roosevelt in a state). Let $M_1, M_2, \ldots, M_K$ represent K candidate models considered for predicting θ. Then the posterior distribution of θ given the data D is

$$\Pr(\theta|D) = \sum_{k=1}^{K} \Pr(\theta|M_k, D)\Pr(M_k|D).$$

The posterior distribution of θ is a weighted average of the posterior distribution of θ given each of the candidate models, with the weights determined by the posterior probabilities of those models.

For the situation of the 1936 election, the data from 1936 have no information about the models, because the models are designed to address errors from undercoverage and nonresponse. Thus, for this application, the posterior distribution $\Pr(M_k \mid D)$ is set equal to a prior distribution $\Pr(M_k)$. The prior distribution may be set subjectively, or may be partially informed by previous results. We caution against relying exclusively on previous results when setting prior probabilities, because they do not capture changes in conditions that the various weighting models are designed to overcome. For the LD polls, the ratio weighting method worked well for the 1928 data and did no harm for the 1932 data, but conditions had changed by 1936.

For unweighted estimates, the posterior distribution for a proportion p, using the noninformative Jeffreys prior, is a Beta $(x+1/2,\ n-x+1/2)$ distribution, which for large sample sizes is approximately normal. For poststratified estimators, ignoring the finite population correction and the small-sample correction in Little (1993), the posterior distribution for the proportion has mean $\hat{p}_{PS} = \sum_{h=1}^{H} W_h \hat{p}_h$ and variance $\sum_{h=1}^{H} W_h^2 s_h^2 / n_h$, where there are H poststrata, W_h is the population proportion in poststratum h, and s_h^2 and n_h are the variance and sample size within poststratum h. The Bayesian estimate $\hat{p}_h$ is the mean of a Beta $(x_h+1/2,\ n_h-x_h+1/2)$ distribution, which for large sample sizes is approximately the sample proportion in the poststratum.

Thus, for election results, the posterior distribution under Bayesian model averaging is approximately a mixture of normal distributions weighted by the prior probabilities of the models. Figure 5 shows the 95 percent highest posterior density (HPD) intervals for the predictions from each state, assuming that the unweighted model and the model in equation (1) each have prior probability ½. This assumes that the analyst has no information about which weighting model might be preferred; of course, different prior distributions could be used that would give different results and we present this model for illustration purposes only. For this analysis, states with a total of 158 electoral votes are predicted for

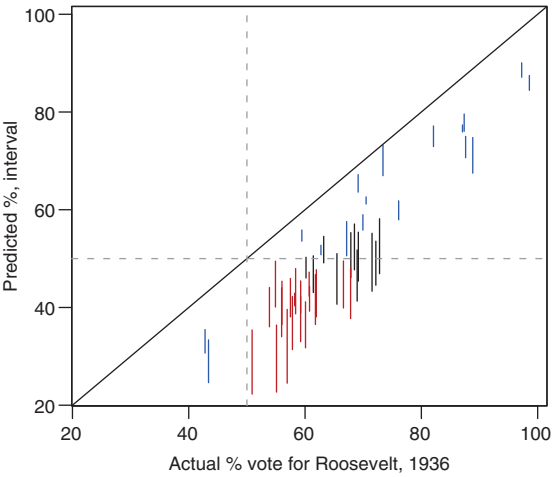


Figure 5: 95 percent highest posterior density intervals for Bayesian model averaging method. Note: The blue lines represent states where Roosevelt won and the interval is entirely above 50 percent, or Roosevelt lost and the interval is entirely below 50 percent; the red lines represent states where Roosevelt won and the interval is entirely below 50 percent; and the black lines represent states where Roosevelt won and the interval includes 50 percent.

Roosevelt (the lower bound of the HPD interval is above 50 percent), 255 electoral votes are predicted for Landon (the upper bound of the HPD interval is below 50 percent), and 118 electoral votes are indeterminate (the HPD interval includes 50 percent).

While the Bayesian model averaging does increase the uncertainty about the estimated winner in this case, it still fails to account for most of the nonsampling errors because the models are similar. At least part of the problem with the 1936 situation is the lack of other data such as demographic data that might facilitate constructing other realistic models. These types of data are available in today's polls and could be used to construct a more diverse set of models. Additional weighting models or projections might also be considered, which make use of external information. The result of using a Bayesian model averaging rather than relying on a single weighting model will often be to move the predicted probability of winning closer to 50 percent because of the additional uncertainty in the individual state results.

It might be argued that analyses such as model averaging are unnecessary because poststratification should in theory do no harm. However, for election forecasting, many different models are used that have different determinations of likely voters, party affiliation, and nonresponse adjustments. In addition, pollsters often make a bias/variance tradeoff decision when deciding how fine the

weighting categories should be. In the 2016 US presidential election, for example, *The Los Angeles Times* used small weighting cells for adjusting its poll (deciding that the potential bias reduction was more important than the variance inflation), and that decision was criticized by some of the other pollsters (see Cohn 2016b and Lauter 2016).

Weighting adjustments can in fact increase bias if the poststratification totals or the classification information from the survey have errors. Willcox (1931) discussed potential objections to the representativeness of the sample collected in the 1930 LD poll about retention or repeal of Prohibition, and concluded: "So far as I can judge they do not suffice to destroy the value or seriously to impeach the representative character" of the 1930 sample (Willcox 1931: p. 249). The Literary Digest (1932a) quoted statements from Willcox's radio address on the topic, and viewed his statements as an endorsement of their methods. The LD provided Willcox with a tabulation of ballots mailed from cities with more than 5000 inhabitants and those mailed from other areas. The ballots represented 4.8 percent of the "city folk" but only 3.1 percent of the "country folk," but Willcox (1931: p. 246) argued that postmarks may have been inaccurate if "country folk" mailed their ballots from a city. Poststratifying to Census counts of the rural and urban population using the postmarks could have been considered as an alternative weighting of the data, but we do not know if it would have reduced bias. If results differed with a weighting based on postmarks, this would represent another source of model uncertainty and we argue that this uncertainty should be included in the uncertainty of the estimates.

Spiegelhalter and Riesch (2011) argued that uncertainty about models should be incorporated into risk assessments. They also discussed uncertainty that can be attributed to inadequacies of all the models. For the 1936 LD poll, it is possible to construct a weighting model, using heroic assumptions, that in hindsight gives the correct result. However, none of the weighting models that reasonably could have been considered at the time come close to predicting the electoral or popular vote margin that Roosevelt experienced in 1936. Even the best model is seriously flawed. Although demographic information might have been used in the weighting had it been available, many of the relationships between demographic characteristics and voting behavior that had held at the beginning of the 1920s no longer held by 1940 so that demographic weighting would also have likely produced inaccurate results. A major lesson of the 1936 LD poll is that models that held before a period of transition may not be useful during or after that period. Spiegelhalter and Riesch (2011: p. 4744) held that it is preferable to "have an external perspective on the adequacy and robustness of the whole modelling endeavour, rather than relying on within-model calculations of uncertainties,

which are inevitably contingent on yet more assumptions that may turn out to be misguided."

6 Revisiting the *Literary Digest* Poll of 1936

2016 marked the centennial of the first *Literary Digest* poll in 1916. The first poll in 1916 was of five states; subsequent polls sampled most or all states and predicted the correct winner, even though the popular vote margins could be inaccurate. By 1936, after four correct election predictions, the editors were confident about their methodology and launched the 1936 poll by writing: "By the time the Poll is completed … [w]e should know exactly how Roosevelt and Landon will fare. For this Poll of 1936 is being conducted in precisely the same manner as the many others which have proved nearly 100 per cent. correct every time in pointing out the winner long before he won" (Literary Digest 1936b: p. 8). Their contemporaries had mixed opinions; often the opinions depended on whether the poll predicted that a preferred candidate would win. In 1924, Democrats attacked the poll (which predicted Coolidge would win) on the grounds that three times as many ballots were sent to Republicans as to Democrats. They said (New York Times 1924): "Straws may show which way the wind is blowing, but not when it is a whirlwind."

In this paper, we examined the results that would have been obtained by the *Literary Digest* polls in 1928, 1932, and 1936 had they used a simple ratio weighting for each state. Had they weighted the data using the method in Section 3, they would have predicted Roosevelt to win the 1936 election. The *Digest* still would have been wildly incorrect on the margin of Roosevelt's win and on the popular vote. This partial reduction of bias by adjustment is typical of what can be expected. But some have attributed the magazine's sale and subsequent demise at least in part to the failure of the 1936 poll (New York Times 1937), and it is possible that these outcomes might have been different had they predicted the correct winner as they had in previous elections, even if the percentage predictions were off.

Lusinchi (2012: p. 45) concluded: "For social scientists who rely on sample surveys as their source of data, the lesson from the *Digest* is that any poll or survey that has a low response rate is probably biased." Lusinchi then noted the similarity to today's polls that use large internet panels, which have large numbers of respondents but are self-selected.

In our opinion, however, the *Literary Digest* poll of 1936 was not as uniformly bad as it has been portrayed. They did many things right, including publishing

details about their methodology and their response rate. The poll was a probability sample – intended to be a census – from the sampling frame constructed from telephone directories and automobile registration lists. It achieved a response rate of 24 percent. The response rate and perhaps even the coverage of the population are much higher than current polls. The *Literary Digest*'s publication of its methodology was also an example that today's polls and aggregators might emulate.

The *Literary Digest* pollsters deserve credit for collecting auxiliary information on the candidate voted for in 1932, which can be used to examine potential bias in the estimates. They also recorded (but did not publish) the postmarks for the returned ballots, and could have used them to look at urban/rural differences in responses, or used county-level information on demographics from the 1930 census to perform an alternative weighting of the data.

The main problem with the 1936 *Literary Digest* poll is that despite knowing about these potential biases, they failed to assess the uncertainty this implied about their estimates. How critical should we be about this when today's surveys typically report a margin of error that assumes that weighting adjustments have removed all of the bias? The result is confidence or posterior prediction intervals that severely understate the uncertainty in the results. Those errors are often ignored in election polls when the winner is predicted correctly, and polling errors are often discussed only when polls fail to predict the winner.

The 2015 United Kingdom general election and the 2016 US presidential election are recent prominent examples of polls failing to predict the winners. In both cases, the "whirlwind" analogy might describe the failures, but the lessons from 1936 suggest that the main problem was the failure to account for the uncertainty of the predictions. The history of polling in US presidential elections shows that the errors in the 2016 polls were typical of the magnitude of errors observed in the past (summaries from 1936 to 2012 are available from the National Council of Public Polls 2017). The main problem is that the uncertainty in the estimates was underestimated, usually accounting for sampling errors only. Sturgis et al. (2016) found the lack of representative samples was the main problem with the 2015 UK general election forecasts and recommended better reporting of the uncertainty of the estimates, but they did not suggest going beyond sampling errors. The 2016 Brexit outcome is somewhat different because the polls, even the aggregates, showed the election too close to call. The surprise associated with the outcome was more because other methods of assessing the outcome such as betting markets suggested Remain was the likely winner (Cohn 2016a). Eighty years after the *Literary Digest* poll of 1936, Harford (2016) commented that "the uncertainties are not going away, so it's not too late to learn."

After every election in which polls err, numerous commentators publish articles about what went wrong with the polls. Gallup's (1938) commentary was

of this type. Deming (1986: p. 319) observed that "Dr. George Gallup remarked in a speech one time (after a fiasco) that he made his prediction in advance of the election. Other people, smarter, made their predictions after the election, explaining how it all happened." In some respects, post hoc explanations view the polling inadequacies as what Deming called a "special cause" attributable to unusual features of that particular election. However, since the outcomes of elections typically differ from the poll estimates by considerably more than sampling error, Deming would argue that this is a system-level, or "common" cause. Our system of assessing the uncertainty of estimates from surveys is inadequate and this needs to be addressed systematically rather than trying to explain what is wrong with a particular outcome. Unfortunately, the main lesson from 1936 has not yet been learned.

Acknowledgments: The authors are grateful to the editor and referees for their helpful comments.

References

Amateur Statistician (1936) "The Literary Digest Poll: Mr. Franklin's Mathematical Analysis Evokes Some Criticism," Letter to the Editor, *The New York Times* (October 31), 18.

Belli, R. F., M. W. Traugott and M. N. Beckmann (2001) "What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies," Journal of Official Statistics, 17:479–498.

Berent, M. K., J. A. Krosnick and A. Lupia (2016) "Measuring Voter Registration and Turnout in Surveys: Do Official Government Records Yield More Accurate Assessments?" Public Opinion Quarterly, 80:597–621.

Bowley, A. L. (1926) "Measurement of the Precision Attained in Sampling," Bulletin of the International Statistical Institute, 22(Supplement to Book I):1–62.

Bryson, M. C. (1976) "The Literary Digest Poll: Making of a Statistical Myth," The American Statistician, 30:184–185.

Cahalan, D. (1989) "Comment: The Digest Poll Rides Again !" Public Opinion Quarterly, 53:129–133.

Campbell, J. E. (2010) "Explaining Politics, Not Polls: Reexamining Macropartisanship with Recalibrated NES Data," Public Opinion Quarterly, 74:616–642.

Cohn, N. (2016a) "Why the Surprise Over 'Brexit'? Don't Blame the Polls," *The New York Times* (June 24). Available at: https://www.nytimes.com/2016/06/25/upshot/why-the-surprise-over-brexit-dont-blame-the-polls.html.

Cohn, N. (2016b) "How One Illinois Man Distorts National Polls," *The New York Times* (October 13), p. A18.

Cornfield, J. (1942) "On Certain Biases in Samples of Human Populations," Journal of the American Statistical Association, 37:63–68.

Crum, W. L. (1933) "On Analytical Interpretation of Straw-Vote Samples," Journal of the American Statistical Association, 28:152–163.

Deming, W. E. (1986) *Out of the Crisis*. Cambridge, MA: MIT Press.
Franklin, F. (1936) "Refiguring the Digest Poll: Returns Otherwise Examined Found to Show Roosevelt Lead," Letter to the Editor, *The New York Times* (October 28), 24.
Gallup, G. (1938) "Government and the Sampling Referendum," Journal of the American Statistical Association, 33:131–142.
Hansen, M. H., W. N. Hurwitz and W. G. Madow (1953) *Sample Survey Methods and Theory* (Vol. 1: Methods and Applications). New York, NY: John Wiley & Sons.
Harford, T. (2016) "When Forecasters Get it Wrong," *Financial Times* (November 26), 45.
Hjort, N. L. and G. Claeskens (2003) "Frequentist Model Average Estimators," Journal of the American Statistical Association, 98:879–899.
Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky (1999) "Bayesian Model Averaging: A Tutorial" (with discussion), Statistical Science, 14:382–417.
Laplace, P. (1814) *Essai Philosophique sur les Probabilités* (no.57), Paris, France: MME VE Courcier, Imprimeur-Libraire pour les Mathématiques, quai des Augustins.
Lauter, D. (2016) "No, One 19-year-old Trump Supporter Probably Isn't Distorting the Polling Averages All By Himself," *The Los Angeles Times* (October 13) [online]. Available at: http://www.latimes.com/politics/la-na-pol-daybreak-poll-questions-20161013-snap-story.html.
Levy, F. F. (1936) "Forecasting the Election: Method Suggested to Correct the Poll Taken by the Literary Digest," Letter to the Editor, *The New York Times* (October 16), 24.
Literary Digest (1928a) "Semi-final Figures in 'The Digest's' Big Poll," Literary Digest, 99 (October 27):10–12.
Literary Digest (1928b) "Final Returns in 'The Digest's' Presidential Poll," Literary Digest, 99 (November 3):5–7.
Literary Digest (1932a) "'Digest' Poll Scrutinized by an Expert," Literary Digest, 114 (October 8): 37–39.
Literary Digest (1932b) "Roosevelt Bags 41 States Out of 48," Literary Digest, 114 (November 5): 8, 9, 44, 46, 47.
Literary Digest (1936a) "'The Digest' Presidential Poll is On!" Literary Digest, 122 (August 22): 3–4.
Literary Digest (1936b) "First Votes in 'Digest's' 1936 Poll," Literary Digest, 122 (September 5): 7–8.
Literary Digest (1936c) "Landon, 1,293,669; Roosevelt, 972,897: Final Returns in 'The Digest's' Poll of Ten Million Voters," Literary Digest, 122 (October 31):5–6.
Literary Digest (1936d) "What Went Wrong with the Polls?" Literary Digest, 122 (November 14): 7–8.
Little, R. J. A. (1993) "Post-Stratification: A Modeler's Perspective," Journal of the American Statistical Association, 88:1001–1012.
Little, B. (2016) "Four of History's Worst Political Predictions," *National Geographic* [online] Available at: http://news.nationalgeographic.com/2016/11/presidential-election-predictions-history/.
Lohr, S. (2010) *Sampling: Design and Analysis* (2nd ed.). Boston, MA: Brooks/Cole.
Lohr, S. and T. Raghunathan (2017) "Combining Survey Data with Other Data Sources," Statistical Science, in press. Available at: http://imstat.org/sts/future_papers.html.
Lusinchi, D. (2012) "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" Social Science History, 36:23–54.
National Council of Public Polls (2017) "Election Results," Available at: http://www.ncpp.org/?q=node/101.

New York Times (1924) "Contest Accuracy of Digest's Poll," *The New York Ti*mes (October 20), 2.
New York Times (1937) "Literary Digest Bought by Shaws," *The New York Times* (June 17), 21.
Neyman, J. (1934) "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, 97:558–625.
Parten, M. (1950) *Surveys, Polls, and Samples*. New York, NY: Harper & Brothers.
Presser, S., M. W. Traugott and S. Traugott (1990) "Vote 'Over' Reporting in Surveys: The Records or the Respondents?" ANES Technical Report Series No. nes010157. Available at: http://electionstudies.org/Library/papers/documents/nes010157.pdf.
Robinson, C. (1932) *Straw Votes*. New York, NY: Columbia University Press.
Spiegelhalter, D. J. and H. Riesch (2011), "Don't Know, Can't Know: Embracing Deeper Uncertainties When Analyzing Risks," Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 369:4730–4750.
Squire, P. (1988) "Why the 1936 *Literary Digest* Poll Failed," Public Opinion Quarterly, 52:125–133.
Sturgis, P., N. Baker, M. Callegaro, S. Fisher, J. Green, W. Jennings, J. Kuha, B. Lauderdale and P. Smith (2016) *Report of the Inquiry into the 2015 British General Election Opinion Polls*. London: Market Research Society and British Polling Council. Available at: http://eprints.ncrm.ac.uk/3789/1/Report_final_revised.pdf.
Walton Jr., H., C. V. Gray and L. McLemore (2001) "African American Public Opinion and the Pre-Scientific Polls: The Literary Digest Magazine's Straw-Vote Presidential Polls, 1916–1936," National Political Science Review, 8:221–243.
Willcox, W. F. (1931) "An Attempt to Measure Public Opinion About Repealing the Eighteenth Amendment," Journal of the American Statistical Association, 26:243–261.
Wright, G. C. (1993), "Errors in Measuring Vote Choice in the National Election Studies, 1952–88," American Journal of Political Science, 37:291–316.