Smartphone Polling for the German Federal Election 2017

Election Forecasting Project

Moritz Hemmerlein & Alexander Sacharow

12 April, 2017

Abstract

Using online surveys to forecast election outcomes imposes severe challenges to pollsters. Non-representative samples and likely-voter bias skew gathered information and require adequate statistical adjustment. This paper proposes a research design to employ different methods on data from mobile phone app users surveyed on their vote intention in the upcoming German federal election.

Contents

1 Introduction

together

The digitalisation is challenging the way polling was done for many decades. Calling people on their land-line phones has become difficult as response rates dropped and households equipped with land-line phones are getting less and less (Skibba 2016). In response, polling institutions have resorted to other methods for polling ranging from face-to-face interviews to mobile phone calling. However, these methods either face similar difficulties to ensure representativeness or are too expensive for regular polling. In order to tackle these obstacles, pollsters are increasingly using online polls, which are cheap and fast, but can be highly non-representative (e.g. W. Wang et al. 2015).

The eventual aim of pollsters in online as well as traditional polling is to collect sample data that reflects the view of a population of interest. The major difference between both methods is that for various reasons online polls cannot ensure representativeness before the actual poll takes place. For instance, respondents of online surveys are more likely to be from certain demographic groups or share a particular political background depending on the website or app where the survey is conducted. However, such non-representative polls can be statistically adjusted to match the demographic composition of the population.

Additionally, online election polls, like traditional ones, face another problem. Election forecasters are naturally not only interested in the population as such, but in the population of actual voters. By the time a poll is made representative in demographic terms, it is still in question whether it reflects the group of people who actually cast their ballots. This, however, is crucial in order to make an accurate prediction. Traditional polling tries to account for this using likely voter models and could perform fairly well (Gallup 2010, Keeter, Igielnik, and Weisel (2016)). Online surveys will also have to be adjusted to actual voting population in order to provide accurate predictions.

In this paper, we want to propose a research design to explore and analyse how different approaches to adjust online polls perform. Our first approach will be a two step procedure where the polling data is first made representative of the population and then likely voter methods are used to resemble the probable voters population. Our second approach attempts to combine both steps into one by adjusting the online polls directly to exit polls and voting statistics from previous elections. For this, we work with individual level data from mobile-phone app users who were surveyed on their vote intention in the German federal election 2017. As the ultimate election will only be after the end of research, we use other forecasts of the 2017 federal election as a benchmark.

The structure of the paper is as follows: Section two will survey the literature on non-representative polls and approaches to employ such data to forecast elections. Subsequently, we present our data and discuss possible problems with it. Afterwards, we discuss the methodology we want to use for adjusting our data at hand.

2 Related Literature

Alexander

Traditional polling and in particular election polling has relied heavily on telephone surveys for the last decades. To ensure representativeness, the standard was randomized digit dialing (RDD). The selection of random respondents was intended to eliminate the sample bias of the survey. However, for several reasons this approach has become unreliable. First, response rates have declined heavily (Keeter et al. 2006; Holbrook, Krosnick, and Pfent 2007). The Pew Research Center (2012) reported that in the U.S. response rates dropped down to 9% in 2012, compared to 36% in 1997. This is fostered by technical changes that, for instance, make it possible to identify the caller before taking the call and increased the likelihood of people not answering survey requests. Second, more and more people do not get land-lines telephones after moving to new places or just give them up as mobile phone and other means of communication have increasingly become popular. As a result, random polls are often exposed to non-response bias mitigating the probabilistic approach of RDD. Hence, classical representative polling is becoming less reliable, a trend that will rather continue than cease. Unsurprisingly, lacking representativeness of surveys has been identified as a core reason for recent election polling failures, e.g. in the UK General elections 2015 (Mellon and Prosser 2015).

Can non-representative online polls fix this problem? Since the famously failure of the Literature Digest poll in the 1936 U.S. presidential election, pollsters have been sceptical of non-representative polling (Squire 1988, Goel, Obeng, and Rothschild (2017)) and this scepticism is still widespread. Yeager et al. (2011), for example, argue that phone surveys are still more accurate than online polls. However, they base their argument on simple correction approaches for non-probability sampled online polls.[1]

W. Wang et al. (2015) in contrast are much more optimistic about the possibilities of non-representative polling. They used polling results from XBox users, which were highly unrepresentative of the population, to forecast the 2012 U.S. presidential elections. By employing a sophisticated multi-stage approach to post-stratify and calibrate the data they were able to generate accurate forecasts of the elections. However, their methodology is mainly suited for large data sets which have sufficient observations for the combined sub-groups (strata) in the sample. Goel, Obeng, and Rothschild (2017) showed that smaller non-representative online surveys can also be accurate. They conducted polls on Amazon Mechanical Turk and a mobile phone app and achieved a level of accuracy sufficient for most practical applications. These works shows that online survey can be used in a meaningful way, if appropriate methods are used to stratify and calibrate the non-representative data. Still, how such methods perform for the case of election forecasting remains contested.

[1] Goel, Obeng, and Rothschild (2017 Footnote 2) where able to get more accurate results using the basic approach, but in a different way.

3 Data and Potential Biases

Moritz

3.1 Europulse Survey

In this paper, we are using data from Dalia Research, an online polling firm which is conducting market and opinion research exclusively through smartphones. To ensure to collect data from a broad variety of target populations, Dalia is using a diverse set of app and website categories such as sports, news, entertainment or games. To control participants answer the survey seriously, an algorithm analyses the consistency and the response behaviour and computes a “trust score” to every respondent. Dalia praises its methodology as distinctively accounting for potential biases such as interviewer effect, social desirability bias or interviewer data entry errors. (Dalia Research 2016)

Our forecasting project utilizes data of Dalia Research’s Europulse Survey which is conducted quarterly in all EU countries. The survey consists of seven waves, but for this project we only use two waves of the survey from December 2016 and March 2017. The first wave is freely available on Kaggle, the second wave was provided to us directly by Dalia Research.[2] Each wave consists of about 11000 individuals, of which roughly 1900 were from Germany which is the fraction of respondents we will focus on in the following analysis. The data is already pre-stratified by Dalia Research based on micro census data for age and gender.

The Europulse data is not particularly collected for election forecasting purposes but contains data on a variety of questions such as online behaviour, media consumption and personal views on political and societal development in the European Union and the respondent’s country of origin. Moreover, the survey contains information on the respondent’s personal background, demographic data and his or her financial situation. These data can be utilized in order to improve the representativeness of the survey through weighting. This will be explained more detailed in the methodology section.

The election-related variables collected in the Europulse survey are similar to traditional vote intention polling questions. First of all, the respondents are asked if and for which party they will vote in the upcoming election and for which party they voted in the previous election. Moreover, they are asked to rank political parties and describe the degree of certainty to cast a ballot for a particular party.

[2] The data for the first wave was collected between 5th and 15th of December 2016. The second wave was conducted between 13th to 27th of March 2017.

3.2 Potential Sources of Bias

As with all surveys, the methodology of Dalia Research has several sources of potential biases. In the case of the Europulse survey they are primarily from flawed measurement and representation of the surveyed population, as listed in figure 1 (Groves et al. 2009).

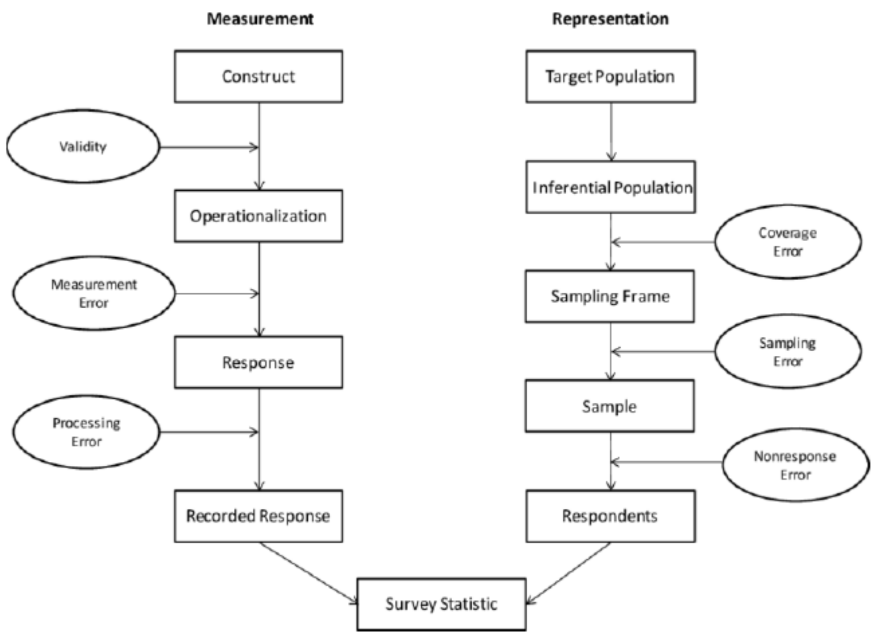
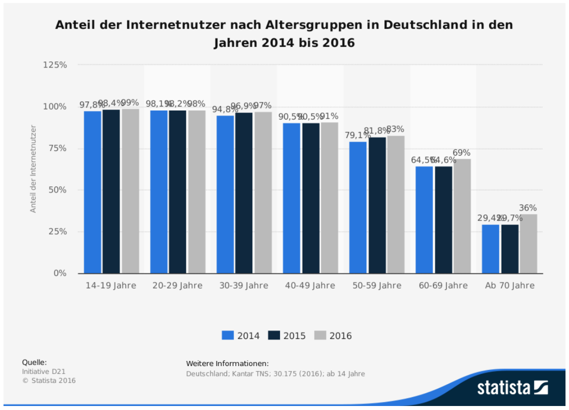


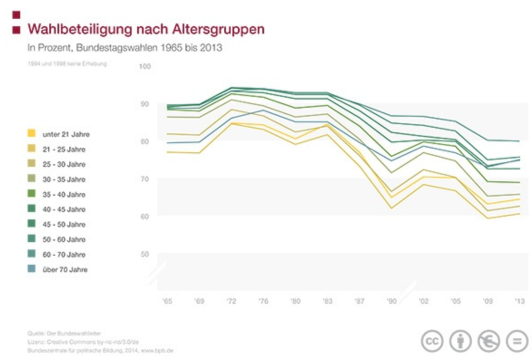
Figure 1: Potential sources of survey error

With regard to measuring vote intention for a particular party, the Europulse survey uses a similar approach as traditional surveys: it asks respondents directly which party they intent to vote for at the upcoming election. Whether such questions measure correctly the actual voting behaviour at the election day is questionable but the approach does not differ from other polling methods with regards to validity. However, online surveys, such as Europulse, can reasonably claim to avoid some sources of measurement error such as social desirability bias or interviewer bias. Since such surveys are often anonymous, the social pressure on the respondent is presumably neglectable, and interaction with the interviewer does not bias the response.

Regarding representativeness, the Europulse survey has some limitations in comparison to face-to-face or RDD interviews. First of all, Europulse' framework does not select participants at random but offers visitors of certain websites or app-users the opportunity to participate in the survey. Hence, this approach carries the risk of self-selection. Moreover, representativeness would require that the Europulse survey in principle should be available to the entire population. However, not everyone is using smart phones and even if they are, they might not use the applications Dalia targets for its surveys. The users Dalia Research actually tends to reach are likely to be much younger and technology oriented than the general population would be (see figure). This is in



(a) Internet usage by age



(b) Voter turnout by age

Figure 2: Internet usage and turnout by age

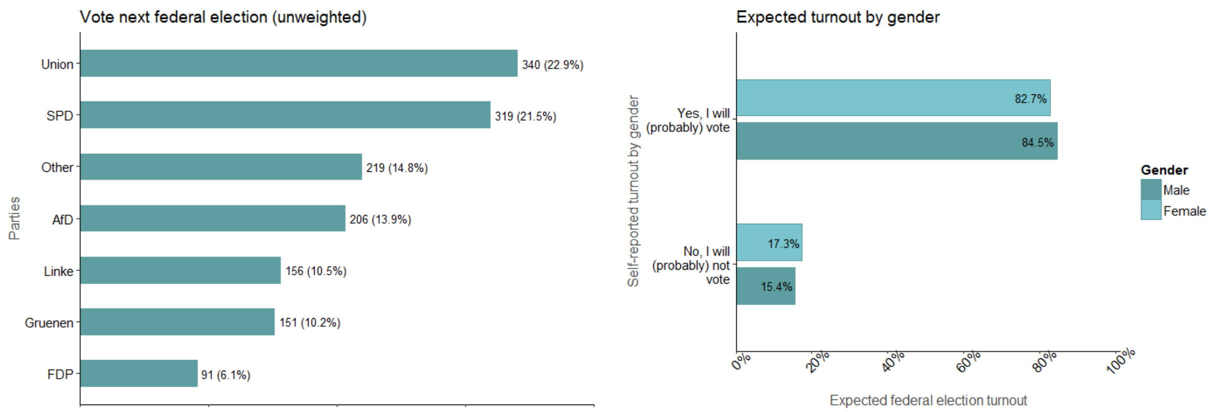
particular a problem for election polling, as older voters are systematically underrepresented in such online methodologies.

Dalia Research tries to account for these problems. First, they try to increase the diversity of their respondents by presenting their surveys on various apps and platforms targeting different user groups. Second, they pre- and post-stratify their data. Pre-stratification is done on the basis of age and gender, using self-reported demographic information. For each age and gender strata they target a certain number of respondents so that their sample resembles roughly the German population. In a second step, they use data from the German Census and compute weights for combined cluster of age, gender, education and whether the respondent lives in an urban or rural area. These weights can finally be used to post-stratify the sample to match the demographic composition of the population more closely. Third, the selection bias is slightly decreased by the fact that respondents are not informed that they are answering an election related survey before they actually start it and in general Dalia Research has high completion rates, hence participants are only rarely dropping out after they started a survey.

Despite these efforts, it is questionable whether the data can be regarded as representative. Printing a simple frequency table of the vote intent at the next general election and using self-reported voting intent to estimate the turnout shows results significantly diverging from other forecasts (see www.wahlumfragen.org).[3]

The large divergence from can be due to a variety of flaws that bias Europulse. First of all, even if the sample represents the German population in terms of age groups and gender, it can be question whether the old female or male users are truly representative of their age group. Moreover, not revealing the survey content might sort out political motivated respondents, but interest in political matters is likely correlated to general willingness to respond to surveys. Hence, in order to utilize

[3] The raw vote intentions are pre-stratified by age and gender, but not adjusted with the weights Dalia Research computed.



(a) (b)

Figure 3: Vote intent and self-reported turnout (Europulse December 2016)

the data to election forecasting further weighting and post-stratification will be necessary.

3.3 Data for Post-stratification and Weighting

The data used for post-stratification and weighting come from several sources. To adjust our sample to demographic composition of the German population we use data of the German Zensus 2011. The data is freely available and we obtained combined frequencies for characteristics such as age, gender, education, employment status and confession. The Census claims to reflect the actual demographic distribution of the German population and hence is suitable in order to post-stratify our sample to demographic criteria.

Furthermore, we collected data from election polls conducted by Forschungsgruppe Wahlen e.V., Infratest dimap and from the official German election statistics (Neu 2013). While the latter contains only information on turnout and the demographic dimensions age and gender, Forschungsgruppe Wahlen e.V. has also issued election results and turnout along groups with different education, different employment status and confession. Such data is the closest estimate of the actual voting population and their voting behaviour across demographic groups and across the spectrum of political parties we can get.

Finally, in order to set our forecasts into context we will use polling data from other institutes. For this we plan to get the respective information from wahlrecht.de.

4 Methodology

We want to employ two approaches to generate election forecasts from our data. In the first approach, the survey data is adjusted in a two step procedure: First, the representativeness of survey data to the general population is increased by finding appropiate weights. Then, a likely voter model is used to obtain an election forecast. Our second approach does this in one step: The survey data will be directly adjusted to the likely voter population based on exit poll data from the German federal election in 2013.

4.1 Raking and post-stratifiation

Alexander

In order to follow both procedures, we need methods to compute weights for different strata of the survey sample aiming at increased representativeness. A classical way to get these weights is *raking*. With raking weights are assigned to each respondent in order to match the marginal distribution of characteristics in the 'true' population. For example, if we know the distribution of education level and employment status in the population we can compute a weight for each so that the weighted survey sample has the same distribution of the characteristics as the 'true' population. Basically, raking estimates a joint distributions for each combination of characteristics.

In formal terms, we can describe each combination of an individual and a characteristic with $x_{i,j} \in \{0,1\}$ where i stands for the individual and j for the characteristic. To illustrate, in the case of the characteristic gender (j): If $x_{i,j} = 1$ the individual i is female and if $x_{i,j} = 0$ the individual i is not female. Hence, each characteristic is modeled with a binary variable. c_j expresses the prevalence of a characteristic in the population and raking estimates the weights w_i such that:

$$c_j = \frac{\sum_{i=1}^{n} w_i x_{i,j}}{\sum_{i=1}^{n} w_i} \forall j \tag{1}$$

The weights can then be used to compute the raking estimates for each observation in the survey y:

$$\hat{y}^{rake} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

However, raking is not ideal. It is often used if only marginal probabilities are available. This is also the case for some of our data. However, if the joint distribution of characteristics is known we can use *post-stratification* instead. In contrast to raking, post-stratification takes the distribution of combinations of characteristics in the population into account. In other words: It sub-divides the population into groups according to their characteristics. Formally, the post-stratification estimator can be expressed as follows (Goel, Obeng, and Rothschild 2017, Lohr (2010, 142f.)):

$$\hat{y}^{post} = \frac{\sum_{j=1}^{J} N_J \hat{y}_j}{\sum_{j=1}^{J} N_J} \text{ where } \hat{y}_j = \frac{1}{n_J} \sum_{i=1}^{n_J} y_i, \tag{3}$$

where $\hat{y}^{post}$ is the estimator of y in the strata j, N_J the size of the j-th strata in the population and n_J the size of the j-th strata in the sample. From the equation it is obvious that post-stratification is only feasible if we have at least one observation for each strata ($n_J > 0$). But even for small strata sizes in the sample ($n_J < 30$) the standard errors will be large. This is a serious issue as the number of strata grows rapidly with each characteristic included. If we use for instance two gender categories and four age groups we have eight strata, if we add past vote (7 categories) we already have 54. As a result, we might only have a few individuals for each strata or even none. This has severe consequences: For example, estimating the population frequency of the strata female, old (60+) and FDP voter in 2013 with only few respondents in the sample will come with a large error.

Furthermore, our effort to post-stratification comes with another problem: Publically available data often only includes the distribution of pairs of characteristics or, in rare cases, triples (e.g. age, gender and past vote). Yet, the ideal would be to have the frequencies of strata of all variables that is correlated to the voting behaviour of a person. These would be certainly gender and age but also employment status, income, education and religion.[4] Knowing this distribution would enable us to fine tune our forecast. However, as they are not available we will mainly concentrate on combination of characteristics where we have the joint distribtuions. These are in particular age, gender and past votes (from exit polls) as well as combinations of age, gender, education, religion and/or employment from the census.

The prior mentioned problem of empthy strata can be address in two ways: By aggregating the data to fill strata, which is sometimes criticised as ad-hoc, or with *model-based post-stratification*. In the model-based post-stratification approach the estimates for each strata are not based on the average in the strata as in the equation above, but the result of a multinomial logistic regression. In order to arrive at this regression results, demographic variables in the sample can be used. If we decide to implement this, we will orient ourself at the work of Goel, Obeng, and Rothschild (2017).[5] However, the implementation ultimately depends on the availability of the strata sizes in the true population. This is the major problem for our (second) application, as the exit polls do not publish the extensive contingency tables which reveal sub-group sizes in their voter sample.

4.2 Likely voter models

Moritz

By the time we post-stratified and weighted our sample to resemble the population (first approach), we have not yet accounted for the problem of likely voter bias. Surveys tend to either include voters who say they will vote but eventually do not do so, or include voters who say they will not vote but finally cast a ballot (Bernstein, Anita, and Montjoy 2001). Since this behaviour of individuals has

[4]Formally speaking, we need to know the number of individuals in each strata N_j

[5]They are developing an r package for this (postr) and might be willing to share their r script (as they already anounced to make it public).

turned out to be correlated across party preferences it is likely to bias our forecast (Keeter, Igielnik, and Weisel 2016)

Two families of methods are proposed by Keeter, Igielnik, and Weisel (2016) to account for likely voter bias - *deterministic* and *propabilistic* methods. The former was initially proposed in Perry (1960) and Perry (1979) and tries to identify criteria that describe a voter as a likely voter on a scale from 0-7. Using a distinct set of questions such as "How likely are you to vote at the next national election?" respondents are assigned points on the likely voter scale and are eventually ranked. Using an estimate of the upcoming election turnout, a cut-of criteria determines which rank on the likely voter scale is needed to be included in the sample measuring the actual voting population.

The downside of such methods is that people below the threshold value of the scale have no affect at all on the actual forecast. This is addressed by probabilistic methods. Such an approach estimates on the basis of a set of predictor variables like demographic characteristics, partisanship or ideology a probability of a person to cast a ballot. By using this probability even people with a very low likelihood of voting affect the actual estimate to a certain degree. The downside of such methods is that it requires data on turnout and the predictor variables of the past election. Moreover, such models assume that turnout is time-stable for the population groups across past and future elections.(Keeter, Igielnik, and Weisel 2016)

Applying such methods on our data, we are basically capable of using deterministic as well as probabilistic identification of likely voters. However, it is easier to use deterministic criteria since they do not require data on voting behaviour in the past election. On the downside, we would need an estimate for the turnout in the federal election. With probabilistic models we are not capable of estimating the probability for the sub-groups using for instance logistic regression since we do not have individual data from past elections. However, we could assign probabilities utilizing the turnout data of the sub-groups we obtained from exit polls. In our paper we will reflect on both options and compare their performance with regard to an actual forecast.

References

Bernstein, Robert, Chadha Anita, and Robert Montjoy. 2001. "Overreporting Voting: Why It Happens and Why It Matters." *Public Opinion Quaterly* 65: 22–44.

Dalia Research. 2016. "Dalia Research Methodology." https://daliaresearch.com/wp-content/uploads/2016/08/Methodology-PDF-1.pdf.

Gallup. 2010. "Understanding Gallup's Likely Voter Models." http://www.gallup.com/poll/143372/understanding-gallup-likely-voter-models.aspx?version=print.

Goel, Sharad, Adam Obeng, and David Rothschild. 2017. "Online, Opt-in Surveys: Fast and Cheap, but Art They Accurate?" Working Paper.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2nd ed. New Jersey: John Wiley & Sons Inc.

Holbrook, Allyson L., Jon A. Krosnick, and Alison Pfent. 2007. "The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms." In *Advances in Telephone Survey Methodology*, edited by James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, 499–528. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9780470173404.ch23.

Keeter, Scott, Ruth Igielnik, and Rachel Weisel. 2016. "Can Likely Voter Models Be Improved? Evidence from the 2014 U.S. House Elections." Pew Research Center.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National Rdd Telephone Survey." *The Public Opinion Quarterly* 70 (5): 759–79. http://www.jstor.org/stable/4124225.

Lohr, Sharon. 2010. *Sampling: Design and Analysis*. Brooks/Cole.

Mellon, Jonathan, and Chris Prosser. 2015. "Investigating the Great British Polling Miss: Evidence from the British Election Study." *SSRN Electronic Journal*. doi:10.2139/ssrn.2631165.

Neu, Viola. 2013. "Bundestagswahl in Deutschland Am 22. September 2013." Konrad Adenauer Stiftung. http://www.kas.de/upload/dokumente/2013/09/Anhang_gesamt_neu.pdf.

Perry, Paul. 1960. "Election Survey Procedures of the Gallup Poll." *Public Opinion Quaterly* 24: 531–42.

———. 1979. "Certain Problems in Election Survey Methodology." *Public Opinion Quaterly* 43: 312–25.

Pew Research Center. 2012. "Assessing the Representativeness of Public Opinion Surveys." Pew Research Center. http://www.people-press.org/files/legacy-pdf/Assessing%20the%20Representativeness%20of%20Public%20Opinion%20Surveys.pdf.

Skibba, Ramin. 2016. "The Polling Crisis: How to Tell What People Really Think." *Nature* 538 (7625): 304–6. doi:10.1038/538304a.

Squire, Peverill. 1988. "Why the 1936 Literary Digest Poll Failed." *The Public Opinion Quarterly* 52 (1): 125–33. http://www.jstor.org/stable/2749114.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91. doi:10.1016/j.ijforecast.2014.06.001.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang. 2011.

"Comparing the Accuracy of Rdd Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75 (4): 709–47. doi:10.1093/poq/nfr020.