Public Opinion Quarterly, Vol. 81, No. 3, Fall 2017, pp. 714–735

USING WIKIPEDIA TO PREDICT ELECTION OUTCOMES
ONLINE BEHAVIOR AS A PREDICTOR OF VOTING

BENJAMIN K. SMITH*
ABEL GUSTAFSON

Abstract This study seeks to improve election forecasting by supplementing polling data with online information-seeking behavior trends as an indicator of public opinion. Aggregate trends of demonstrations of interest or engagement have been shown to accurately predict behavior trends and reflect public opinion. Further, because traditional poll-based predictions are inherently undermined by self-reporting biases and the intention-behavior disconnect, we can expect that information-seeking trends on widely used social media—as an autonomous and unobtrusive indicator of relative levels of public opinion—can help correct for some of this error and explain unique, additional variance in election results. We advance the literature by using data from Wikipedia pageviews along with polling data in a synthesized model based on the results of the 2008, 2010, and 2012 US Senate general elections. Results show that Wikipedia pageviews data significantly add to the ability of poll- and fundamentals-based projections to predict election results up to 28 weeks prior to Election Day, and benefit predictions most at those early points, when poll-based predictions are weakest.

There is a global fascination with predicting the results of elections before they happen. The reasons for this are multifaceted: political pundits use election predictions as fodder for the 24-hour news cycle (Rosenstiel 2005), political strategists use them to direct their resources (Jamieson 2009), and academics use them to better understand the electorate (e.g., Wlezien and Erikson 1996; Panagopoulos 2009). This interest has only increased over the past decade, alongside a sharp increase in the number of pre-election polls (Linzer 2013).

Benjamin K. Smith and Abel Gustafson are doctoral students in the Department of Communication at the University of California, Santa Barbara, CA, USA. The authors would like to thank Joel Gustafson for his assistance with data collection, and Andrew Flanagin for his guidance in conceptualizing and revising this study. *Address correspondence to Benjamin K. Smith, Department of Communication, 4005 SS&MS Building, University of California at Santa Barbara, Santa Barbara, CA 93106–4020, USA; e-mail: bsmith@umail.ucsb.edu.

doi:10.1093/poq/nfx007 Advance Access publication May 6, 2017

This study argues that poll-based election forecasts may be improved, especially during early parts of the election campaign, by incorporating a measure of online information-seeking behavior trends (Wikipedia pageviews). Specifically, we posit that web traffic on candidates' Wikipedia pages is strongly correlated with their relative vote share, can account for some of the error seen in early election polling, and may significantly improve prediction models. To test these hypotheses, we collect data about candidates' Wikipedia pageviews during the 200 days leading up to the 2008, 2010, and 2012 Senate elections, along with the results of almost 2,000 pre-election polls conducted during those same time spans. Utilizing polling data, fundamentals data, and Wikipedia pageviews data in a synthesized model, we show that election opponents' proportion of Wikipedia pageviews were significantly related to vote share in all three sets of elections, and that Wikipedia usage can account for a significant amount of the variance in vote shares that is left unexplained by public opinion polling data or fundamentals data.

Indicators of Public Opinion and Predictors of Behavior

Modern election forecasts inform predictions with diverse indicators of public opinion and other variables known to correlate with election results. This section reviews the use of these diverse predictors in election forecasting, and comments on the emerging line of research that investigates correlations between social behavior trends and election results.

POLLING AND CONTEXTUAL VARIABLES

The most common (and best) indicator of public opinion and predictor of future voting behavior is the traditional public opinion polling of a representative sample of the electorate. However, polling data are subject to multiple sources of error, including sampling, self-reporting biases, and the intention-behavior disconnect (Clausen 1968; Sheeran 2002; Wlezien and Erikson 2002). To correct for this error and to improve accuracy, diverse alternative variables are used to contribute information about public opinion and future voting behavior.

Many forecasts use contextual and historical variables shown to be consistent correlates of voting behavior. For example, the proxy model (Lewis-Beck and Tien 2012) uses the National Business Index—a measure of consumer sentiment. Erikson and Wleizien (2008) use presidential approval ratings, economic trends, and candidates' prior occupations. The time-for-change model (Abramowitz 2012) incorporates gross domestic product (GDP) trends, presidential approval rating, incumbency status, and the polarization of public opinion. Although, for example, economic trends may have a tenuous causal relationship with any individual person's voting behavior, they are still strongly and explicably correlated with aggregate trends of voting behavior. This reliable correlation makes them particularly valuable to election forecasters.

BEHAVIORAL INDICATORS OF FUTURE BEHAVIOR

Public opinion can also be assessed—and voting behavior predicted—by observing existing behavioral patterns in the public. For example, betting markets—where individuals bet on candidates' performance—were considered reliable indicators of election results as far back as the 1800s (Rhode and Strumpf 2004). In modern election forecasts, they are still used as a barometer of public opinion—and are significant predictors of election results (Wolfers and Zitzewitz 2004; Erikson and Wlezien 2008).

Measures that quantify the public conversation surrounding a topic can also indicate public interest or engagement (as proxies for favorable public opinion) and thereby predict behavior. For example, one study found that the number of times a book was mentioned in blogs across the web predicted its performance in Amazon's sales rank (Gruhl et al. 2005). Applying this principle to election forecasting, Véronis (2007) counted the number of times the press mentioned each candidate, and found that each candidate's tally was a better predictor of his/her election results than the polls. Similarly, Williams and Gulati (2008) demonstrated the same finding with the number of "likes" given to a candidate's Facebook page. Tumasjan et al. (2010) found that in the 2009 German parliamentary elections, the simple tally of candidate or party mentions on Twitter accurately predicted how they fared, respectively. According to this growing cluster of research, the quantity of media mentions is an indirect measure of popularity, and thereby of public opinion.

In a well-reasoned critique, Gayo-Avello, Metaxas, and Mustafaraj (2011; see also Metaxas, Mustafaraj, and Gayo-Avello [2011]) discuss the limitations inherent to predicting election outcomes with media mentions, citing concerns of ambiguous sentiment, sample representativeness, and weak linkage between predictor and outcome. We agree that despite the demonstrated correlations, the "predictive" ability of Twitter and news media mentions should be interpreted with caution, because the "behavior" captured by mentioning a candidate's name does not have an intuitive linkage to voting behavior. However, it should be noted that—similar to the use of economic trends—weak causal linkage at the individual level does not negate the practical usefulness of variables that, in the aggregate, have a strong (and explicable) correlation with voting behavior. In the following section, we discuss ways that these limitations can be mitigated and how valuable benefits can be leveraged.

Improving The Use of Behavior Trends to Predict Elections

In this section, we first argue that the use of behavior trends in forecasting can be made somewhat more defensibile by choosing a more valid behavior trend as a predictor. More importantly, we then advance the literature by proposing that social behavior trends should *not* be used as stand-alone predictors of election results, but rather as *complementary* to the polls in synthesized models, so they can reduce error inherent to poll-based predictions and thus explain unique variance.

WIKIPEDIA: IMPROVING THE PREDICTOR

Recent research has sought to mitigate the limitations stated above by choosing a more defensible behavioral variable: online information-seeking trends (Moat et al. 2014). Information-seeking is a common and influential part of the process of political participation, especially in the weeks leading up to an election (Lau and Redlawsk 2006), and at least 36 percent of Americans regularly get their campaign information from the Internet (Rosenstiel and Mitchell 2012). One of the leading drivers of information-seeking behavior is convenience, and often the most convenient choice is Wikipedia, the online encyclopedia (Head and Eisenberg 2010; Connaway, Dickey, and Radford 2011). For example, any Google search for the name of a US state- or national-level political candidate will return that candidate's Wikipedia page on the first page of search results. Wikipedia is one of the most popular and recognizable sites in the world, with its number of unique monthly viewers placing in the top 15 web-wide (comScore 2014; Quantcast 2014). Its massive user base represents a diverse portion of the voting public, with significant user-ship across all age, ethnic, and socioeconomic groups (Zickuhr and Rainie 2011).

Due to Wikipedia's ubiquity, prevalence, and prominence as an information-seeking tool, Wikipedia usage data are a good indicator of public information-seeking trends. Information-seeking patterns, as a reflection of public interest, engagement, and opinion about a topic, have been used to predict diverse public behavior trends. A recent study touted in *Nature*'s online *Scientific Reports* demonstrated the ability of Wikipedia pageviews to anticipate stock market trends (Alanyali, Moat, and Preis 2013). Similarly, Wikipedia pageviews predict movies' performance at the box office (Mestyán, Yasseri, and Kertész 2013). Advancing the election forecasting literature, Yasseri and Bright (2013) used both Wikipedia pageviews and Google search trends to predict the winners of UK and Iranian elections. These studies find consistent support for their assumption that online information-seeking trends are strongly and reliably correlated with future behavior, including voting (Yasseri and Bright 2015).

This research assumes that higher quantities of Wikipedia pageviews are indicative of more positive public interest or opinion. This assumption is reasonable, as meta-analyses of information-seeking research show that people seek out opinion- and behavior-congruent information more than incongruent information (Hart et al. 2009), and follow patterns of selective exposure—especially regarding political topics (for better or worse; see Bennett and Iyengar [2008]). Thus, while the pageviews data are certain to contain a proportion of information-seekers who are antagonistic, it is reasonable to expect that the vast majority will be those who are supportive or at least positively curious—supporting the claim that Wikipedia pageviews are a valid indicator of public interest and opinion in the aggregate. Further, we must remember that research has repeatedly corroborated this premise, despite the unavoidable noise in the Wikipedia pageviews data and its indirect nature as a measure of public opinion. Therefore, we expect to find that:

H1: The number of pageviews a candidate's Wikipedia page receives in comparison with the competing candidate in their race will be positively associated with the proportion of votes they receive in the general election.

IMPROVING THE MODEL AND EXPLAINING UNIQUE VARIANCE

The use of behavior trends to predict election outcomes can be improved not only by optimizing which predictors are used, but also by optimizing how they are used. Therefore, in this section, we first explain the limitations of forecasting elections with a single predictor, and then summarize the benefits of synthesized models—arguing that they can mitigate the limitations of behavioral predictors while retaining the ability of behavioral measures to correct for error inherent to poll-based predictions.

Simple models: In general, election forecasters and researchers should be wary of using big data barometers of social behavior as *replacements* for poll-based models. Many of the studies we have surveyed use social behavior predictors in "simple models," which simply correlate one predictor variable (Tweets, "likes," pageviews, etc.) with the election results. There are three central limitations to such models. First, although each single predictor explains a statistically significant amount of variance, none approach the levels of the traditional, rigorous election forecasts. Therefore, while interesting, these correlations are not practically useful. Second, because each predictor is tested in separate studies, the variance explained may or may not be unique variance that is beyond what is already captured by traditional poll-based models. Simply put, the extant research has not yet shown that these demonstrated correlations can actually improve upon an already rigorous forecast. Third, behavioral indicators of public opinion (even information-seeking trends) are theoretically and empirically far inferior to explicit, self-reported intentions of future behavior, such as those represented in polling data (Gayo-Avello, Metaxas, and Mustafaraj 2011). Thus, to rely on behavior trends *instead of* the polls is to eliminate the most reliable, valid, and accurate indicator of voting behavior. The use of social behavior trends in simple models invokes all of the critiques and limitations discussed above. However, we argue that the limitations can be mitigated, and some benefits leveraged, by utilizing online information-seeking trends to *complement* polling data in a synthesized model.

Synthesized models: Many of the most sophisticated and accurate modern election forecasting models combine contextual variables and/or aggregate behavior trends along with the polling data (Linzer 2013; Lewis-Beck and Dassonneville 2015a,b). For example, Erikson and Wlezien (2008) integrate contextual variables such as economic trends, previous election outcomes, approval ratings, and candidates' occupational history (termed "fundamentals") *along with* the polling data. Rothschild (2015) combines betting markets with fundamentals data and polling data.

The benefit of such synthesized models is that they capture a set of well-rounded information that together can minimize the effect of the error of its component parts (Graefe et al. 2014). Similarly, Rothschild (2015) advocates for his synthesized model by specifically arguing for the value of diversified sources of predictive information. Combining forecasts of disparate types (polling, expert opinion, betting markets, etc.) has been shown to reduce error by more than 50 percent, compared to the individual forecast of each component (Graefe et al. 2014). This, then, is the true value of behavioral and contextual variables in election forecasting. While they are inferior to polling data as stand-alone predictors, they can reduce significant amounts of error when used alongside polling data.

Reducing error by adding Wikipedia: The addition of behavioral predictors increases accuracy because they account for some of the error that is left unexplained by polling data alone. In this section, we argue that the addition of a behavior trend measure (Wikipedia pageviews) as a predictor can correct for some leading sources of polling error, thus explaining unique variance.

Polling error: The error inherent to basing predictions on self-reported intentions can arise from multiple causes. An individual may simply change their opinion or lose interest before Election Day, and expressed opinions can differ widely from private opinions when social pressures are at play (Asch 1951; Edelman and Mitofsky 1990). Poll respondents can also be influenced by a social desirability bias to not publicly express ignorance or apathy toward civic engagement (Clausen 1968), leading people to express an opinion where none exists (Price 1992). Even legitimate and strong intentions are not a perfect predictor of voting behavior. Ajzen, Brown, and Carvajal (2004) report the effect of hypothetical bias—that there is a disconnect between the self-reported anticipated behavior and actual behavior when placed in a real-life situation.

Correcting for error: The use of behavior trends, such as online information-seeking, directly combats the intention-behavior gap by measuring already-existent behavior, rather than stated intentions of it. Further, the unobtrusive, autonomous, and self-motivated nature of these measures circumvents the biases of self-report, resulting in further reduction of error. This pattern is also evidenced at the individual level, where decades of research have shown that while stated intentions (analogous to their aggregate counterpart, the polls) are the best single predictor of behavior, the error inherent in the self-reporting of intentions and the intention-behavior gap can be significantly reduced by incorporating measures of prior or existing behavior (Fishbein and Ajzen 2011).

Previous research has demonstrated the correlation between Wikipedia pageviews data and election results (Yasseri and Bright 2013, 2015). However, the extant literature has only used trend-level behavioral indicators of public opinion in simple models, which—as we discussed—limits their substantive contribution. If behavior trends like online information-seeking were included alongside polling data in a synthesized model, this would advance the literature with a more informative and practically valuable test of the utility of Wikipedia as a predictor of election results. Further, the autonomous nature of online information-seeking trends such as Wikipedia pageviews and the unobtrusive nature of data collection could naturally correct for some error inherent in poll-based predictions in a synthesized model. Therefore, we expect that not only will we find that Wikipedia pageviews are correlated with election results (H1), but also that they can explain a significant amount of *unique* variance, such that:

H2: The number of pageviews a candidate's Wikipedia page receives in comparison with the competing candidate in their race will predict unique variance in the proportion of votes they receive in the general election, beyond the variance explained by the polling results alone.

Methods

This study seeks to determine whether Wikipedia pageviews can add to the ability of horse-race polls to predict election outcomes, and to demonstrate the relationship between Wikipedia pageviews and electoral outcomes. For this preliminary investigation, we chose to focus on the 104 Senate general election races that occurred in November 2008, 2010, and 2012. This total includes five special elections and excludes one uncontested election. We did not look at elections prior to 2008, because the relevant Wikipedia data are only available after 2007. All of our models start 200 days before the election (dbe) of each respective year, due to the precedent of prior research (e.g., Erikson and Wlezien 2008), and because polling data is exponentially less available 200+ dbe.

DEPENDENT VARIABLE—ELECTION RESULTS

The dependent variable is the relative two-party vote share. We operationalize vote share as the fraction of the two-party vote received by the Democrat (or equivalent)—that is, the number of votes cast in the general election for Candidate 1 divided by the sum of the votes received by both Candidate 1 and Candidate 2. For example, if the Democrat in a race received 100,000 votes and the Republican received 110,000 votes, then the two-party vote share would be $\frac{100{,}000}{100{,}000+110{,}000} = 47.6 \text{ percent}$. Online appendix A contains a detailed discussion of why we chose to use Democrats as the referent and to use the two-party vote share.

Official vote counts were those published by the Office of the Clerk of the US House of Representatives ("Election Statistics, 1920 to Present" 2015), which are compiled from the official sources. For simplicity and congruence with prior research (e.g., Erikson and Wlezien 1999; Rothschild 2015), we chose to analyze only the two candidates in each race who received the highest proportion of the vote in the general election. We did not investigate runoff or primary elections.

DATA COLLECTION AND TRANSFORMATION INTO "VOTE SHARE" STYLE VARIABLES

Polls: We gathered polling data from the *New York Times*'s extensive database of Senate polls,[1] used in their Senate forecast model "Leo." All general election Senate polls for the races in our sample were included. Upon running the Leo model in R, an output file is created that contains the two-party poll share of the Democrat candidate for each of the historical polls. As such, no additional transformations were required.

Wikipedia: Data for Wikipedia pageviews were collected from stats.grok.se, which compiles Wikipedia pageviews via the regular data dumps provided by Wikimedia (http://dumps.wikimedia.org/other/pagecounts-raw/). Data are

[1] According to the *New York Times* website, polls in its database are collected from "Pollster, the Roper Center for Public Opinion Research, the U.S. Officials' Job Approval Ratings project, Polling Report, Gallup, Talking Points Memo, The Argo Journal, Real Clear Politics and FiveThirtyEight." All data for the Leo model, including the historical polling results used, are made publicly available at https://github.com/TheUpshot/leo-senate-model.

available from the creation of a page until the latest data dump, which occurred roughly once a month.[2] If one or both of the main candidates in a race did not have a Wikipedia page for the entirety of the election cycle, we did not collect pageview statistics for that race. Finally, in races where one of the pages was not created until after data collection began, views for both candidates are only counted from the start of the newest page. In some instances, the Wikipedia pageviews for all pages are missing (the longest such stretch occurs in 2008 between 115 and 96 dbe).[3] In these instances, we follow the precedent set by others (e.g., Rothschild 2009), and perform a linear interpolation to fill in the missing data.

The raw Wikipedia pageviews data were transformed into a "vote share"–style variable by taking the natural log (ln) of the pageviews for the Democratic candidate and dividing it by the sum of the log-transformed pageviews for both the Democratic candidate and the Republican candidate. For example, in the 2008 Arkansas Senate election between incumbent Democrat Mark Pryor and Republican Rebekah Kennedy, Pryor had 560 page views 30 dbe, while Kennedy had 21. The view share for the Democrat would then be $\frac{\ln(560)}{\ln\left(560\right)+\ln(30)} = \frac{2.75}{2.75+1.45} = .65$, or 65 percent.

This example illustrates the value of the ln transformation, because without it, the view share for Pryor would have been $\frac{560}{(560)+(30)} = .95$, or 95 percent, a value alarmingly close to the maximum of 1. Such highly skewed results would be frequent in the dataset. Figure 1 shows the distribution of the share of pageviews variable 30 dbe using both the untransformed and ln transformed data. In the raw pageviews method (top), the distribution is almost perfectly flat, whereas the distribution of the ln pageviews method (bottom) appears close to normal. The Shapiro-Wilk test shows that when the raw pageviews are used, the null hypothesis of normality is rejected, $S\text{-}W(84) = .960$, $p = .010$. The ln transformation corrects for this issue (e.g., Cohen et al. 2003, chap. 6), such that when the pageviews have been ln transformed, the Shapiro-Wilk test suggests that normality is a reasonable assumption: $S\text{-}W(84) = .975$, $p = .098$. For a detailed conversation about when it may or may not be appropriate to ln transform, and further discussion of the implications of non-normality as it applies to the assumptions of OLS estimation, see online appendix A, and see the descriptives section of our results.

[2] In early 2015, the method of compiling pageviews data changed, and the dataset is now available on an hourly basis. See https://wikitech.wikimedia.org/wiki/Analytics/Data/Pagecounts-all-sites#Availability.

[3] In 2008, pageviews data are missing for all pages on the following days before the election (in reverse chronological order): 13–14, 96–115, 126, and 155–56 dbe. In 2010, data are missing 115–18, 120, 127, and 129 dbe. In 2012, data are only missing 190 dbe.

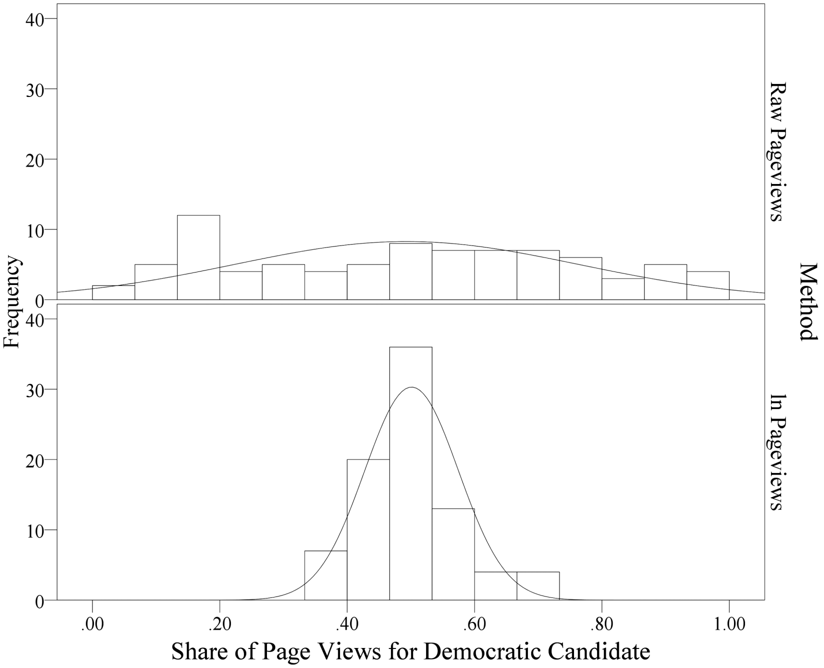

Figure 1. Comparison of Distributions Using Raw Pageviews vs. the Natural Log (ln) of Pageviews. Histograms show the distribution of the share of pageviews for the Democratic candidate 30 days before the election (n = 84), alternatively calculated using the raw pageviews or the natural log (ln) of the pageviews. Bin intervals are equal in both the top and bottom histogram.

TRANSFORMATION OF RAW DATA INTO PROJECTIONS

Polls: While an intuitive way to determine the ability of Wikipedia data to improve the predictive ability of polling results might be to take the raw polling results and use them as the poll-based prediction of the final election outcome, the need to transform raw polling data into a forecast, or projection, "is conclusive in the literature" (Rothschild 2015, 952). This mitigates well-documented polling biases and improves the overall accuracy of predictions. Additionally, projections tend to outperform more simple methods like raw polling averages (Pasek 2015).

The following methods for developing our projections model are derived from Rothschild (2009). The first step is creating a "snapshot" that—for each of the 200 dbe—estimates the two-party vote share if the election were held that day. This snapshot is simply the results of the most recent poll conducted in that race (if more than two polls were conducted that day, then they are pooled), oriented as vote share of the Democratic party candidate (see online appendix A). We then take an additional step, as advocated by Rothschild (2009), of debiasing the polls, using the linear trend of all polls up to that day[4] as the snapshot (for more discussion on this method, see online appendix A). This step reduces the day-to-day volatility in projections, creating a more stable forecast (Rothschild 2009, 2015).

The second step in this process is to create the actual projection of the final outcome. This is done by regressing the observed vote share of the Democratic candidate on the debiased snapshot for each day prior to the election: $V_r = a + bS_r + e_r$, where V is the observed vote share of the Democratic candidate, r is a given race (i.e., state and year), and S is the snapshot (see Rothschild 2015, 957). The parameters a and b are calibrated separately for each dbe, and then used to create a daily estimate of the final vote share for each dbe (t): $\hat{V}_r = a_t + b_t S_r$. This two-step method has been used in many other studies to great effect (see, e.g., Lewis-Beck and Dassonneville 2015b).

Wikipedia pageviews: While significant research exists on transforming raw polling data into Election Day projections, few studies use Wikipedia pageviews to predict elections, and we are unaware of any that do so in the context of US elections. Thus, the "best" method for transforming raw pageviews into a projection is still unknown. We argue above that information-seeking trends are indicative of aggregate-level voter opinions, so we presume it reasonable for this preliminary analysis to treat the pageviews data similar to the polling data. As such, we calculate the daily estimate of the final view share based on Wikipedia pageviews with the same method as we did for calculating the daily estimate of the final vote share based on polls. The snapshot is simply the share of ln pageviews for the Democratic candidate, calculated and debiased as described above. Similarly, the snapshot is then transformed into a projection, by regressing the observed vote share of the Democratic candidate on the debiased snapshot for each day prior to the election: $V_r = a + bS_r + e_r$, where V is the observed vote share of the Democratic candidate, r is a given race (i.e., state and year), and S is the snapshot (see Rothschild 2015, 957). As with the polling projections, the parameters a and b are calibrated separately for each dbe, and then used to create a daily estimate of the final vote share for each day before the election (t): $\hat{V}_r = a_t + b_t S_r$.

FUNDAMENTALS

Many of the most robust election forecasting models incorporate "fundamentals" data into their poll-based projections, creating what are known as

[4] In cases where polling data are not available from 200 dbe, the linear trend begins on the day the first poll is released in the race.

"synthesized" models, as discussed above. We employ the same practice, to ensure that we test whether Wikipedia pageviews data can improve upon the most rigorous election projection possible, relying upon the fundamentals model of Hummel and Rothschild (2014). The model includes 20 variables and seven different types of data: (a) presidential approval, (b) incumbency, (c) past election results, (d) economic indicators, (e) state ideology, (f) senator ideology, and (g) candidate characteristics. Across the 578 Senate elections conducted between 1976 and 2012,[5] this model was able to account for 71.8 percent of variance in election outcomes. We re-created this model using the exact specifications provided by Hummel and Rothschild (2014, 127), with our resulting fundamentals projection accounting for 69.2 percent of the variance (adj. $R^2 = .689$, $n = 104$), with a root mean square error of 4.60.

SYNTHESIZED PROJECTION MODEL

Hypothesis 2 predicts that the number of pageviews a candidate's Wikipedia page receives in comparison with the competing candidate in their race will predict unique variance in the proportion of votes they receive in the general election, beyond the variance explained by the polling results alone. To that end, we develop a synthesized projection model, using hierarchical multiple regression, with fundamentals and polling data entered at step one, and Wikipedia pageviews data entered at step two. This approach allows us to measure the extent to which Wikipedia decreases the error in the projections (measured as ΔRMSE, or the change in the root mean square error) and increases the variance explained (measured as ΔR^2), and test whether the ΔR^2 is significant.

A key challenge arises when dealing with races where there are data missing from one source (pageviews or polls). Only 59 percent of races have polling data 28 weeks before the election, rising to 83 percent 14 weeks before the election, and 94 percent two weeks before the election. Similarly, Wikipedia pageview projections are available in 69 percent of races 28 weeks before the election, rising to 77 percent 14 weeks before the election, and finally 82 percent two weeks before the election. We could either ignore races when one data source is missing at a particular time, or we could impute the data. Because it is clear that the data are not missing at random (e.g., landslide races are not polled often), we opted to take the latter approach.

To do this, we use the data imputation strategy of Rothschild (2015), whose synthesized model was based on fundamentals, polling data, and betting markets. Specifically, for any day that one data source is missing, the projection

[5] This total excludes special elections, elections where the state has a nonpartisan legislature, and elections in which a third-party candidate received more than 10 percent of the vote. The authors note, however, that including these races does not "does not hurt our forecasted probabilities of winning" (Hummel and Rothschild 2014, 134).

based on the fundamental data takes their place. To ensure that this does not violate the model assumptions of multiple regression, the maximum VIF value at step two is reported, as well as the Durbin-Watson statistic. Importantly, we only impute missing data in the synthesized projection model, which combines the projections of all three types of data. In all other instances, missing data are left as missing.

FORECAST OR CALIBRATION: THE NATURE OF OUR MODELS

It is important to clarify the type of "forecast" that we are conducting in this study. Most forecasting models calibrate the regression parameters used to transform the snapshot into forecasts of expected two-party vote share by only using a subset of the available data, and then test the model using the parameters on out-of-sample data. For example, Rothschild's (2015) model for forecasting the 2012 Senate elections was calibrated using data from the 2004, 2006, 2008, and 2010 senatorial elections. This calibration procedure helps avoid overfitting the model based on the idiosyncrasies of a single election cycle, but requires a large amount of prior data for performing the calibration, which, in the case of Wikipedia, simply isn't available.

Because Wikipedia data are available from only three election cycles, we do not have a large enough data set to accurately calibrate our model and then use the model to forecast an out-of-sample election. As such, the results of this study are best interpreted not as a true "forecast" of each election cycle, but more as an indication that Wikipedia may be able to forecast future elections, based on the post-election model fit demonstrated in our results.

Results

The results are split into three sections. First, we report descriptive statistics about the Wikipedia pageviews data, and the pageviews share variable created from these statistics. We then report the results of the individual election projections created using fundamentals data, polling data, and Wikipedia pageviews data. We conclude by presenting the results of the full synthesized model.

DESCRIPTIVE STATISTICS

Table 1 contains descriptive statistics for transformed Wikipedia pageviews, as well as the corresponding descriptive statistics for the share of pageviews variable. The same statistics for raw Wikipedia pageviews can be found in online appendix B. The raw pageviews totals on any given day are highly leptokurtic, and tend to be skewed to the right. The natural log transformation described in the methods section corrects for this issue, substantively reducing both the skew and kurtosis of the data. Additionally, looking at the Shapiro-Wilk test of

Table 1. Descriptive Statistics for Log-Transformed Wikipedia Pageviews and Democratic Candidates' Share of Log-Transformed Pageviews

	ln(Pageviews)								Share of ln(Pageviews)				
	n	*M*	*Mdn*	*SD*	Range	Skew	Kurtosis	*S-W*	*n*	*M*	*SD*	Range	*S-W*
28 Weeks	146	4.95	4.96	1.39	8.82	–0.1	0.8	.99	73	50.70	11.11	73.78	.93**
26 Weeks	150	5.05	5.05	1.45	9.07	0.0	0.7	.99	75	51.73	11.36	73.86	.99
24 Weeks	152	5.13	5.04	1.50	9.16	0.3	1.1	.97*	76	51.28	10.71	62.96	.96*
22 Weeks	152	5.06	5.13	1.41	8.47	–0.2	1.0	.99	76	51.06	11.28	75.77	.98
20 Weeks	156	4.87	4.90	1.38	7.97	0.0	0.5	.99	78	50.25	10.62	64.44	.96*
18 Weeks	158	5.04	5.00	1.29	7.39	0.3	0.4	.97	79	50.35	9.50	47.60	.99
16 Weeks	158	5.06	5.12	1.33	7.05	0.0	–0.1	.99	79	50.35	9.80	59.01	.99
14 Weeks	160	5.21	5.19	1.30	7.99	0.4	1.1	.98	80	50.53	8.74	49.09	.98
12 Weeks	164	5.30	5.22	1.41	7.39	0.3	0.3	.99	82	49.48	8.83	47.89	.97
10 Weeks	164	5.49	5.41	1.42	8.48	0.2	0.5	.99	82	49.07	8.61	48.73	.98
8 Weeks	166	5.48	5.40	1.28	6.44	0.2	0.2	.99	83	49.73	7.57	33.32	.99
6 Weeks	166	5.83	5.71	1.38	8.88	0.5	1.0	.97	83	50.04	7.29	40.64	.97
4 Weeks	170	6.06	5.99	1.35	8.18	0.4	1.1	.99	85	49.93	6.74	33.69	.98
2 Weeks	170	6.36	6.26	1.30	7.71	0.5	0.7	.99	85	50.27	5.54	29.07	.99

NOTE.—*S-W* = Shapiro-Wilk test statistic. The null hypothesis of the Shapiro-Wilk test of normality is that the data are normally distributed. This test is widely considered the most powerful test of normality, under a wide variety of conditions (Yap and Sim 2011). If the test is significant (especially at $p < .001$), it indicates a significant departure from normality. The significance tests are two-tailed. Descriptive statistics for raw pageviews (i.e., non-log-transformed pageviews) and for Democratic candidates' share of raw pageviews, see online appendix B.

$^{*}p < .05$; $^{**}p < .01$.

normality, it is clear that the ln transformation sufficiently normalizes the data at most points in the election cycle.

The same general trend is true when the raw pageviews and the ln transformed pageviews are transformed into a share of pageviews statistic (as described in the methods section). While the share of pageviews variable based on raw pageviews is not as non-normally distributed as the raw pageviews themselves, the Shapiro-Wilk test of normality is still significant at $p < .05$ at all points except two and four weeks before the election. In contrast, the share of pageviews variable based on ln transformed pageviews is much more consistently normally distributed. The Shapiro-Wilk test of normality is only significant at three points in the election cycle (28 weeks, 24 weeks, and 20 weeks).

INDIVIDUAL PROJECTIONS OF SENATE ELECTIONS

Hypothesis 1 predicted that Senate candidates' share of pageviews will be positively associated with their general election vote share. We began by determining the coefficient of determination (R^2) for the projections at each day before the election, then tested the significance of the relationship between the share of Wikipedia pageviews and vote share. The relationship was significant at all 200 dbe, $p < .001$. The minimum variance explained is 36 percent, with a maximum variance explained of 49 percent roughly one month before the election. This provides strong support for hypothesis 1. We observe a dip in the variance explained approximately 140 dbe (or 20 weeks before the election), possibly caused by a post–primary season lull in candidates' views, distorting the data. However, even at the models' weakest point, the relationship between the expected two-party vote share and the observed two-party vote share was statistically significant.

Figure 2 contains the results of the individual projections of Senate elections, as measured by the root mean square error (RMSE). This statistic is calculated as $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$, and was selected based on its use in prior election forecasting literature (e.g., Lewis-Beck and Dassonneville 2015b), and not only because it gives an indication of the general quality of the resultant model, but also because it gives more weight to large errors (as opposed to mean absolute error, and other measures of absolute error, which tend to obfuscate large misses in projections; see Chai and Draxler [2014]). For completeness, figure 2 includes the results for both the share Wikipedia pageviews based on the ln-transformed pageviews and the raw pageviews—in addition to the results of the poll-based projections and the fundamentals-based projections.[6]

[6] Figure 2 also includes the RMSE for the full synthesized model, as reported in the next section. It is worth noting that the synthesized model has a higher RMSE, but that this is largely due to the fact that the synthesized model is creating projections for all 104 elections, whereas the poll-based projection only includes between 62 and 99 of the elections.

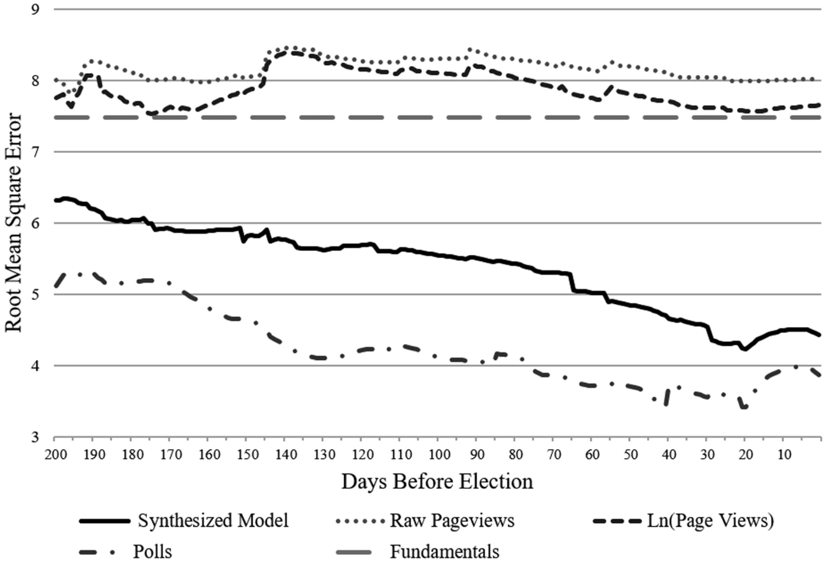

Figure 2. Absolute Errors for Each Projection Type. Absolute errors are measured by root mean square error (RMSE). For the two pageviews-based projections, *n* ranges from 72 to 84 observations per day. For the poll-based projection, *n* ranges from 62 to 99 observations per day. For the fundamentals-based projection and the synthetic model projections, *n* = 104.

The poll-based projection is the most accurate of the individual variable projections, although it is only able to provide a projection for 59.6 percent of the races. Roughly one month before the election, by which time 94 percent of the races have polling data, the RMSE for the poll-based projection decreases to 3.42, the lowest RMSE for any of the projections. As expected, the ln-transformed Wikipedia pageview-based projection does not perform as well as the poll-based projection, and as expected does not perform as well as the fundamentals data (which has an RMSE of 7.49). The RMSE for the ln(Pageviews)-based projection hovers around 7.6 for most of the election cycle. It does, however, outperform the projection based on the raw Wikipedia pageviews at all times in the election cycle.

SYNTHESIZED MODEL COMBINING FUNDAMENTAL, POLLING, AND WIKIPEDIA PROJECTIONS

Hypothesis 2 predicted that the number of pageviews a candidate's Wikipedia page receives in comparison with that of the competing candidate in their race will predict unique variance in the proportion of votes they receive in the general election, beyond the variance explained by the polling results alone.

To test this hypothesis, we ran a two-step hierarchical multiple regression, entering the fundamentals-based projection and the poll-based projection at step one, and entering the ln-transformed Wikipedia pageviews-based projection at step two. As discussed above, this analysis attempts to model all 104 in-sample elections. To achieve this goal, we replace missing data in the poll-based projections and the pageviews-based projections with the fundamentals-based projection. The results of the hierarchical multiple regression can be found in table 2. To ensure that there are no issues with autocorrelation caused by the inclusion of the pageviews projection in the model, or multicollinearity caused by our approach toward imputing missing data, the Durbin-Watson statistic and the maximum variance inflation factor (as well as unstandardized regression coefficients) are reported in online appendix C.

At each two-week interval, the ln-transformed Wikipedia pageviews-based projection was able to explain a significant amount of variance unaccounted for by the poll- and fundamentals-based projections, at $p < .05$, with the sole exception of week 20 ($p < .10$). Additionally, a substantive reduction in the RMSE was seen at each of the two-week intervals checked. Two-thirds of the time, the reduction in RMSE is greater than 0.2, and the reduction is greater than 0.1 at all points except 20 weeks before the election. Across all elections, the average number of pageviews 20 weeks before the election is lower than at any other point in the election cycle, which may explain the decrease in variance explained, and the smaller than normal reduction in RMSE. Despite the marginal improvement shown at 20 weeks before the election, we feel confident in saying that hypothesis 2 is supported.

Discussion

This study sought to determine whether election predictions could be significantly improved by including measures of online information-seeking behavior via a prominent social media venue. We predicted that as an indicator of public interest, the relative number of Wikipedia pageviews for each Senate candidate would be positively associated with their proportion of votes in the 2008, 2010, and 2012 Senate general elections. Strong support for this hypothesis emerged across all three elections.

Additionally, we predicted that pageviews would account for a significant portion of unique variance in election outcomes, that is, variance not previously accounted for by a poll-based synthetic model. This hypothesis was supported, such that the addition of Wikipedia usage data accounted for a significant portion of unique variance across the entire 28-week sampling frame, with the exception of week 20, contributing the most where the polling data are at their weakest point. A potential explanation for the non-significant improvement at week 20 could be the simultaneous occurrence of primaries, which may change the composition and motivations of the population of online information-seekers.

Table 2. Results of Synthesized Projection Model, Combining Fundamentals, Polls and Wikipedia Pageviews Projections

	Model 1				Model 2						
			Fund.	Polls					Fund.	Polls	Pageviews
	R^2	RMSE	b^*	b^*	R^2	ΔR^2	RMSE	ΔRMSE	b^*	b^*	b^*
28 Weeks	.757	6.55	.409***	.499***	.773	.016*	6.34	–0.22	.320**	.407***	.213*
26 Weeks	.777	6.28	.408***	.519***	.795	.018**	6.02	–0.25	.301**	.437***	.224**
24 Weeks	.787	6.13	.409***	.528***	.803	.016**	5.90	–0.24	.306***	.448***	.215**
22 Weeks	.789	6.10	.383***	.550***	.802	.013*	5.91	–0.20	.296***	.472***	.195*
20 Weeks	.806	5.85	.355***	.587***	.812	.006	5.77	–0.09	.301***	.539***	.124
18 Weeks	.810	5.79	.344***	.600***	.819	.009*	5.65	–0.15	.276**	.539***	.156*
16 Weeks	.810	5.80	.336***	.606***	.823	.013**	5.60	–0.20	.257**	.538***	.181**
14 Weeks	.817	5.69	.336***	.612***	.827	.010*	5.53	–0.16	.268**	.548***	.161*
12 Weeks	.819	5.65	.342***	.609***	.831	.011*	5.47	–0.18	.270***	.549***	.164*
10 Weeks	.828	5.51	.327***	.628***	.841	.013**	5.31	–0.21	.253**	.562***	.174**
8 Weeks	.854	5.08	.299***	.670***	.864	.010**	4.90	–0.18	.240***	.604***	.156**
6 Weeks	.861	4.96	.233***	.728***	.874	.014**	4.72	–0.25	.174**	.645***	.178**
4 Weeks	.883	4.56	.204**	.766***	.893	.010**	4.35	–0.21	.156**	.689***	.157**
2 Weeks	.878	4.65	.233***	.740***	.888	.011**	4.44	–0.21	.188**	.658***	.159**

NOTE.—R^2 = coefficient of determination, or the total variance explained by all variables in model; RMSE = root mean square error, a measure of absolute error in the projection, measured in units of vote share; b^* = the estimated value of the standardized regression coefficient; ΔR^2 = change in the coefficient of determination when adding the Wikipedia pageviews projections; ΔRMSE = change in the root mean square error when adding the Wikipedia pageviews projections. All significance tests are two-tailed, $n = 104$. If at any point poll data or Wikipedia pageviews data do not exist (i.e., no polls have been conducted in that race, or there is one or more candidate that does not have a Wikipedia page), then the projection based on the fundamental data takes their place. Unstandardized regression coefficients and standard errors, as well as results of tests for autocorrelation and multicollinearity, can be found in online appendix C.

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

The results of this study hold important implications for the field: They demonstrate not only that online communicative behaviors can serve as a powerful predictor of electoral outcomes, but also that they can reduce a significant amount of the error left over by an already rigorous forecast. This study also advances the election forecasting literature methodologically by incorporating *both* online behavior patterns and polling data in a synthesized model, resulting in a rigorous model that explains almost 90 percent of the variance in vote share.

The limitations of using social media data to make predictions are well documented, and this study is not immune to them. First, while most of the views, for most candidates, can be logically attributed to individuals within that candidate's state, there is no way to identify the actual location of users participating in these activities.[7] In addition, there is no way to know whether a given viewer is actually eligible to vote. However, using large datasets at the aggregate level helps combat this limitation by capturing general social trends.

Another limitation to this study is our use of just one type of election. It is possible that our results are unique to this particular context and, until replicated, should be interpreted with caution. Additionally, although the methods we used are based squarely in the election forecasting literature, they are relatively basic in comparison to the most sophisticated predictive models, which make use of bootstrapping techniques and/or Monte Carlo simulations, which can increase the robustness of model estimates. In sum, the intriguing findings of this study warrant further investigation, using additional measures of communicative behavior, looking at a larger variety of electoral contexts, using more robust prediction models, and/or using the parameters derived in this study to forecast an out-of-sample election.

Despite the limitations, our findings give a clear indication that social behavior patterns, such as Wikipedia pageviews, can improve upon already-rigorous election forecasts up to 28 weeks prior to the election. Not only does this show the power of communicative behavior patterns to predict future behaviors, but it also supports the argument that a significant portion of error in poll-based predictions can be accounted for by factoring in existing behavior trends as an addition and complement to polling data. This study advances the body of literature that has used online behavior patterns to predict elections by building a synthesized model that uses *both* online behavior and a poll-based model, resulting in significantly better predictions than either approach is capable of separately.

[7] A handful of races were particularly impacted by this fact, an exemplar of which is the Missouri race between Claire McCaskill and Todd Akin in 2012. On August 19 of that year, Akin gave an interview with a local Fox affiliate, in which he said: "If it's a legitimate rape, the female body has ways to try to shut the whole thing down." The next day his daily pageviews soared from an average of 176 to 90,216. While this instance is clearly an outlier, and can be accounted for in future analyses, it highlights the larger impact on these data brought from the inability to determine the relationship between viewers of a candidate's page and their constituent status relative to that candidate.

Supplementary Data

Supplementary data are freely available at *Public Opinion Quarterly* online.

References

Abramowitz, Alan I. 2012. "Forecasting in a Polarized Era: The Time for Change Model and the 2012 Presidential Election." *PS: Political Science & Politics* 45:618–19.

Ajzen, Icek, Thomas C. Brown, and Franklin Carvajal. 2004. "Explaining the Discrepancy between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation." *Personality and Social Psychology Bulletin* 30:1108–21.

Alanyali, Merve, Helen Susannah Moat, and Tobias Preis. 2013. "Quantifying the Relationship Between Financial News and the Stock Market." *Scientific Reports* 3: 3578, doi:10.1038/srep03578. Available at http://www.nature.com/articles/srep03578.

Asch, Solomon E. 1951. "Effects of Group Pressure upon the Modification and Distortion of Judgments." In *Groups, Leadership, and Men: Research in Human Relations*, edited by Harold Guetzkow, 177–90. Oxford: Carnegie Press.

Bennett, W. Lance, and Shanto Iyengar. 2008. "A New Era of Minimal Effects? The Changing Foundations of Political Communication." *Journal of Communication* 58:707–31.

Chai, T., and R. R. Draxler. 2014. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments against Avoiding RMSE in the Literature." *Geoscientific Model Development* 7:1247–50.

Clausen, Aage R. 1968. "Response Validity: Vote Report." *Public Opinion Quarterly* 32:588–606.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Lenoa S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. New York: Routledge.

comScore. 2014. *"comScore Ranks the Top 50 U.S. Digital Media Properties for September 2014."* October 22. Available at http://www.comscore.com/Insights/Market-Rankings/comScore-Ranks-the-Top-50-US-Digital-Media-Properties-for-September-2014.

Connaway, Lynn Sillipigni, Timothy J. Dickey, and Marie L. Radford. 2011. "'If It Is Too Inconvenient I'm Not Going after It': Convenience as a Critical Factor in Information-Seeking Behaviors." *Library & Information Science Research* 33:179–90.

Edelman, Murray, and Warren J. Mitofsky. 1990. "The Effect of the Interviewer's Race in Political Surveys with Multiracial Candidates." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Lancaster, PA, USA.

"Election Statistics, 1920 to Present." 2015. *History, Art & Archives, US House of Representatives*. Accessed November 14, 2015. Available at http://history.house.gov/Institution/Election-Statistics/Election-Statistics/.

Erikson, Robert S., and Christopher Wlezien. 1999. "Presidential Polls as a Time Series: The Case of 1996." *Public Opinion Quarterly* 63(2):163–77.

———. 2008. "Are Political Markets Really Superior to Polls as Election Predictors?" *Public Opinion Quarterly* 72(2):190–215.

Fishbein, Martin, and Icek Ajzen. 2011. *Predicting and Changing Behavior: The Reasoned Action Approach*. New York: Taylor & Francis.

Gayo-Avello, Daniel, Panagiotis Takis Metaxas, and Eni Mustafaraj. 2011. "Limits of Electoral Predictions Using Twitter." Proceedings of the 5th International Conference on Weblogs and Social Media. Menlo Park, CA: AAAI Press. Available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2862/3254.

Graefe, Andreas, J. Scott Armstrong, Randall J. Jones Jr., and Alfred G. Cuzán. 2014. "Combining Forecasts: An Application to Elections." *International Journal of Forecasting* 30:43–54.

Gruhl, Daniel, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2005. "The Predictive Power of Online Chatter." In *Proceedings of the Eleventh ACM SIGKDD*

International Conference on Knowledge Discovery in Data Mining, 78–87. Available at http://dl.acm.org.proxy.library.ucsb.edu:2048/citation.cfm?id=1081883.

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill. 2009. "Feeling Validated versus Being Correct: A Meta-Analysis of Selective Exposure to Information." *Psychological Bulletin* 135:555–88.

Head, Alison J., and Michael B. Eisenberg. 2010. "How Today's College Students Use Wikipedia for Course-Related Research." *First Monday* 15(3). Available at http://firstmonday.org/ojs/index.php/fm/article/view/2830.

Hummel, Patrick, and David Rothschild. 2014. "Fundamental Models for Forecasting Elections at the State Level." *Electoral Studies* 35(September): 123–39.

Jamieson, Kathleen Hall, ed. 2009. *Electing the President, 2008: The Insiders' View.* Philadelphia: University of Pennsylvania Press.

Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns. Cambridge Studies in Public Opinion and Political Psychology*. New York: Cambridge University Press.

Lewis-Beck, Michael S., and Ruth Dassonneville. 2015a. "Forecasting Elections in Europe: Synthetic Models." *Research & Politics* 2:1–11.

———. 2015b. "Comparative Election Forecasting: Further Insights from Synthetic Models." *Electoral Studies* 39(September):275–83.

Lewis-Beck, Michael S., and Charles Tien. 2012. "Election Forecasting for Turbulent Times." *PS: Political Science & Politics* 45:625–29.

Linzer, Drew A. 2013. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* 108:124–34.

Mestyán, Márton, Taha Yasseri, and János Kertész. 2013. "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data." *PloS One* 8:e71226.

Metaxas, Panagiotis Takis, Eni Mustafaraj, and Daniel Gayo-Avello. 2011. "How (Not) to Predict Elections." In *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 165–71. Available at http://ieeexplore.ieee.org.proxy.library.ucsb.edu:2048/xpls/abs_all.jsp?arnumber=6113109.

Moat, Helen Susannah, Tobias Preis, Christopher Y. Olivola, Chengwei Liu, and Nick Chater. 2014. "Using Big Data to Predict Collective Behavior in the Real World." *Behavioral and Brain Sciences* 37:92–93.

Panagopoulos, Costas. 2009. "Campaign Dynamics in Battleground and Nonbattleground States." *Public Opinion Quarterly* 73:119–29.

Pasek, Josh. 2015. "Predicting Elections: Considering Tools to Pool the Polls." *Public Opinion Quarterly* 79:594–619.

Price, Vincent. 1992. *Public Opinion*. Newbury Park, CA: Sage Publications.

Quantcast. 2014. *"Wikipedia.org Traffic and Demographic Statistics by Quantcast."* December. Available at https://www.quantcast.com/wikipedia.org.

Rhode, Paul W., and Koleman S. Strumpf. 2004. "Historical Presidential Betting Markets." *Journal of Economic Perspectives* 18:127–41.

Rosenstiel, Tom. 2005. "Political Polling and the New Media Culture: A Case of More Being Less." *Public Opinion Quarterly* 69:698–715.

Rosenstiel, Tom, and Amy Mitchell. 2012. *"Internet Gains Most as Campaign News Source but Cable TV Still Leads."* Pew Research Center's Project for Excellence in Journalism. Available at http://www.journalism.org/2012/10/25/social-media-doubles-remains-limited/.

Rothschild, David. 2009. "Forecasting Elections Comparing Prediction Markets, Polls, and Their Biases." *Public Opinion Quarterly* 73:895–916.

———. 2015. "Combining Forecasts for Elections: Accurate, Relevant, and Timely." *International Journal of Forecasting* 31:952–64.

Sheeran, Paschal. 2002. "Intention—Behavior Relations: A Conceptual and Empirical Review." *European Review of Social Psychology* 12:1–36.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10:178–85.

Véronis, Jean. 2007. *"Citations Dans La Presse et Résultats Du Premier Tour de La Présidentielle 2007."* Accessed December 15, 2009. Available at http://blog.veronis.fr/2007/05/2007-la-presse-fait-nouveau-mieuxque.html.

Williams, Christine, and Girish Gulati. 2008. "What Is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries." *American Political Science Association*. Available at http://195.130.87.21:8080/dspace/handle/123456789/1021.

Wlezien, Christopher, and Robert S. Erikson. 1996. "Temporal Horizons and Presidential Election Forecasts." *American Politics Research* 24:492–505.

———. 2002. "The Timeline of Presidential Election Campaigns." *Journal of Politics* 64:969–93.

Wolfers, Justin, and Eric Zitzewitz. 2004. *"Prediction Markets."* National Bureau of Economic Research, Working Paper 10504. Available at http://www.nber.org/papers/w10504.

Yap, Bee Wah, and Chiaw Hock Sim. 2011. "Comparisons of Various Types of Normality Tests." *Journal of Statistical Computation and Simulation* 81(12):2141–55.

Yasseri, Taha, and Jonathan Bright. 2013. "Can Electoral Popularity Be Predicted Using Socially Generated Big Data?" arXiv:1312.2818 [Physics], December. Available at http://arxiv.org/abs/1312.2818.

———. 2015. "Predicting Elections from Online Information Flows: Towards Theoretically Informed Models." arXiv:1505.01818 [Physics], May. Available at http://arxiv.org/abs/1505.01818.

Zickuhr, Kathryn, and Lee Rainie. 2011. *"Wikipedia, Past and Present."* Pew Research Center: Internet, Science & Tech, January 13. Available at http://www.pewinternet.org/2011/01/13/wikipedia-past-and-present/.