

# A New Multinomial Accuracy Measure for Polling Bias

**Kai Arzheimer**

*University of Mainz, Dept. of Political Science, Colonel-Kleinmann-Weg 2, Mainz, Germany*  
*e-mail: kai.arzheimer@googlemail.com (corresponding author)*

**Jocelyn Evans**

*University of Leeds, POLIS, Leeds LS2 9JT, United Kingdom*  
*e-mail: j.a.j.evans@leeds.ac.uk*

Edited by R. Michael Alvarez

In this article, we propose a polling accuracy measure for multi-party elections based on a generalization of Martin, Traugott, and Kennedy's two-party predictive accuracy index. Treating polls as random samples of a voting population, we first estimate an intercept only multinomial logit model to provide proportionate odds measures of each party's share of the vote, and thereby both unweighted and weighted averages of these values as a summary index for poll accuracy. We then propose measures for significance testing, and run a series of simulations to assess possible bias from the resulting folded normal distribution across different sample sizes, finding that bias is small even for polls with small samples. We apply our measure to the 2012 French presidential election polls to demonstrate its applicability in tracking overall polling performance across time and polling organizations. Finally, we demonstrate the practical value of our measure by using it as a dependent variable in an explanatory model of polling accuracy, testing the different possible sources of bias in the French data.

## 1 Introduction

Work on pre-election polls forms a vital part of election analysis and forecasting in both academic publications and media coverage. Frederick Mosteller (Mosteller et al. 1949) introduced the notion of accuracy measures to assess polls against actual election results. Those indices are designed for two-party/-candidate races, and cannot easily be applied to multi-party/-candidate elections. An equivalent index has not so far been derived for multi-party elections, limiting the ability of researchers to measure overall polling accuracy in such cases.

Starting from the predictive accuracy measure proposed by Martin, Traugott, and Kennedy (2005), we propose such an accuracy measure,  $B$ , for elections with more than two parties or candidates. First, we derive this index mathematically from an implementation of the multinomial logistic model, including the relevant tests of statistical significance. We then consider how this aggregate measure may be biased, given it is based upon compositional data, and use a simulation to examine the extent of this bias. Finally, we use the  $B$  measure as the dependent variable in an explanatory model of different sources of polling bias, as an illustration of how the measure may be applied empirically.

## 2 A Generalisation of Martin, Traugott, and Kennedy's Measure of Bias

### 2.1 The Martin, Traugott, and Kennedy Approach

Martin, Traugott, and Kennedy (2005) propose a measure,  $A$ , for survey bias that is the natural logarithm of an odds ratio (OR). Illustrating their approach with the competition between the two

---

*Authors' note:* We are grateful to two anonymous reviewers and to the editors of *Political Analysis* for their extremely helpful comments and suggestions. We would also like to thank Gilles Ivaldi (CNRS/University of Nice) for providing data on the French election surveys. Replication materials can be found in Arzheimer and Evans (2013). `surveybias.ado`, a Stata add-on that facilitates the computation of  $B$  and  $B_w$ , is available from the Boston College Statistical Software Components (SSC) archive. Supplementary materials for this article are available on the *Political Analysis* Web site.

major American parties, they define the numerator as the odds of a Republican choice in a given pre-election poll  $r / d$ , where  $r$  is the number of respondents favoring the Republicans and  $d$  the equivalent number of Democrats. The denominator of this ratio is  $R / D$ , where  $R$  and  $D$  are the numbers of Republican/Democratic voters in the election. Consequently,  $r$ ,  $d$ ,  $R$ , and  $D$  can be interpreted as the respective proportions of respondents and voters.

As Martin, Traugott, and Kennedy show,  $A$  is superior to earlier measures of poll accuracy. In its original form, however,  $A$  is restricted to two-party systems,<sup>1</sup> and is therefore inapplicable to the majority of democratic systems. Martin et al. provide a suggestion as to how a three-party index might be adapted from their two-party index, but this only measures the accuracy of the poll for a third party as a fraction of the two main parties' total vote (Martin, Traugott, and Kennedy 2005, 367). It does not provide a single measure across all three parties.

## 2.2 Generalization for the Multi-Party Case

To generalize the Martin, Traugott, and Kennedy approach for a choice between  $k$  parties, we define  $\mathbf{p}$  as a vector of proportions  $p_1, p_2, \dots, p_k$  of respondents who support party  $i$  in a given poll, and  $\mathbf{v}$  as a vector of proportions  $v_1, v_2, \dots, v_k$  of citizens who actually vote for the respective party.<sup>2</sup> Applying this generalized terminology to a two-party race, Martin, Traugott, and Kennedy's measure becomes

$$A = \ln \left( \frac{\frac{p_1}{p_2}}{\frac{v_1}{v_2}} \right). \quad (1)$$

Note that  $p_1, p_2, v_1, v_2$  could be written as a  $2 \times 2$  table:  $\begin{pmatrix} p_1 & p_2 \\ v_1 & v_2 \end{pmatrix}$ .  $A$  is then identical to the log of the familiar OR, a central concept in the analysis of tabular data (Agresti 2002, 44–46). In an application with  $k > 2$  parties, there are  $k - 1$  unique ORs (Agresti 2002, 56). This suggests calculating a series of logged ORs as per-party measures of bias, choosing one arbitrary party as the reference.

There are, however, two problems with this approach. First, the clear interpretation of  $A$  would be lost, since the log OR is a measure of bias *relative to the party used as the reference*. Unless support for the reference party is measured without bias, the results will be misleading. To see why, assume a simple case where support for the reference party  $r$  is overestimated in a given poll, but support for another party  $a$  is measured accurately. It follows that  $p_r > v_r$ , whereas  $p_a = v_a$ . Under these assumptions, the odds ratio for party  $a$  with respect to the reference party is

$$\text{OR}_a = \left( \frac{\frac{p_a}{p_r}}{\frac{v_a}{v_r}} \right) = \left( \frac{p_a v_r}{p_r v_a} \right) = \frac{v_r}{p_r} \Leftrightarrow \text{OR}_a < 1, \quad (2)$$

which results in a negative log OR although support for party  $a$  is measured without bias. In short, a measure based on the (log) OR would be biased proportionate to the reference party's own polling bias. Second, it follows that, because the (log) ORs capture bias relative to the reference party, their sum (and consequently their average) depends on the choice of the reference party, making it impossible to derive a unique measure of overall bias.

We therefore propose a different strategy. Equation (1) can be rewritten as

$$A = \ln \left( \frac{\frac{p_1}{p_2}}{\frac{v_1}{v_2}} \right) = \ln \left( \frac{\frac{p_1}{1-p_1}}{\frac{v_1}{1-v_1}} \right). \quad (3)$$

<sup>1</sup>Or more generally to dichotomous variables with a known distribution in the population, such as gender or registration as a voter.

<sup>2</sup>We can treat (self-declared) nonvoters as an additional party.

Using our terminology and following the suggestions from Martin, Traugott, and Kennedy (2005) as well as Durand (2008), there is a straightforward way to further generalize  $A$  for use in a multi-party competition:

$$A'_i = \ln \left( \frac{\frac{p_i}{\sum_{j=1}^k p_j}}{\frac{v_i}{\sum_{j=1}^k v_j}} \right) \text{ for } j \neq i \quad (4)$$

$$= \ln \left( \frac{\frac{p_i}{1-p_i}}{\frac{v_i}{1-v_i}} \right) = \ln \left( \frac{p_i}{1-p_i} \times \frac{1-v_i}{v_i} \right) \quad (5)$$

$$= \ln \left( \frac{p_i}{1-p_i} \times \frac{1}{O_i} \right), \quad (6)$$

where  $A'_i$  represents bias with respect to party  $i$ , and  $O_i$  are the odds (i.e.,  $v/1-v$ ) of a vote for this party. The measure of any  $A'_i$  is therefore not conditional upon an arbitrary selection of reference party.

$A'_i$  retains the clear interpretation of Martin, Traugott, and Kennedy's original  $A$ : positive values indicate bias in favor of party  $i$ , whereas negative values imply bias against  $i$ . In the (highly unlikely) event that a poll is in perfect agreement with the result of the actual election, all  $A'_i$  are zero.<sup>3</sup>

If one is willing to treat a given sample as fixed and is only interested in describing the extent of bias in that sample, calculation of  $A$  and  $A'_i$  is a trivial algebraic exercise. In most settings, however, a survey sample is treated as a single realization of a random process that under essentially identical conditions could have produced an infinite number of similar but slightly different samples. From this perspective,  $A$  and  $A'_i$  are just estimates for the true systematic bias that results from house effects, social desirability, or real changes in the population after the poll was taken. As a consequence, the question of the precision (standard error) of these estimates is crucial.

To our knowledge, there is no existing derivation of an index functionally equivalent to  $A'_i$  and its standard error to define bias in a multiparty system, and consequently no software package available to calculate any such estimators. It is, however, possible to derive these quantities from the parameters and standard errors of an equivalent Multinomial Logit Model (MNL).

To see how the MNL and  $A'_i$  are related, consider an MNL containing only intercepts, where the probability  $p_i$  of reporting a vote for party  $i$  is constant and does not depend on any covariates. As is common for MNLs, the model is parametrized by treating one party as the reference category to which all comparisons refer. For simplicity's sake, define party 1 as the reference, but this choice is arbitrary. Since the model has no explanatory variables,  $p_i$  can be rewritten (cf. Long 1997, 154) as

$$p_i = \frac{\exp(\beta_i)}{\sum_{j=1}^k \exp(\beta_j)}, \quad (7)$$

where  $\beta_i$  is the respective constant from the intercepts only MNL for party  $i$  and

$$\beta_1 = 0 \quad (\text{the reference category}).$$

If we substitute this parametrization into equation (5), we can calculate  $A'_2$  as

$$A'_2 = \ln \left( \frac{\frac{\exp(\beta_2)}{\sum_{j=1}^k \exp(\beta_j)}}{1 - \frac{\exp(\beta_2)}{\sum_{j=1}^k \exp(\beta_j)}} \times \frac{1-v_2}{v_2} \right). \quad (8)$$

<sup>3</sup>Moreover, like in the case of the original  $A$ , calculating  $\exp(A'_i)$  recovers the OR.

Although redefining  $A'_i$  as a nonlinear function of  $k - 1$  parameters might appear complicated, estimating the  $\beta$ s by maximum likelihood generates asymptotically correct estimates for the standard errors of these coefficients. These estimates can in turn be combined to derive an asymptotic standard error for  $A'_i$ .<sup>4</sup>

Inevitably, use of the MNL may raise concerns over assumptions of independence from irrelevant alternatives (IIA). The notion of independence from IIA has a long and convoluted history in game theory and (social) choice theory (Ray 1973) that harks back to the eighteenth century (McLean 1995). Luce (1959, 9) introduced IIA as a property of individual choices which are stochastic but internally consistent, and derived the logit formula from this property (Train 2009, 34).

Substantively, the IIA assumption states that the odds  $p(a)/p(b)$  of preferring alternative  $a$  over alternative  $b$  are not affected if new alternatives are added to the choice set or existing alternatives are removed. In terms of the statistical model, the corollary of the IIA assumption is a specific error structure. Let  $U_{nj} = V_{nj} + \epsilon_{nj}$  be the utility that decision maker  $n$  derives from choosing alternative  $j$ . Let  $V_{nj}$  be the *representative utility* that depends on observable characteristics of the decision maker and the alternatives, and let  $\epsilon_{nj}$  be some additional unobserved utility component (Train 2009, 14–15). Finally, assume that although the utility itself is unknown to the analyst, the decision maker always chooses the alternative yielding the highest utility.

Then the logit model is obtained by assuming that “each  $\epsilon_{nj}$  is independently, identically distributed [iid] extreme value” (Train 2009, 34). If one alternative is perceived as a substitute for another alternative, their random utility components will be correlated, rendering the assumption of independent errors untenable and the logit specification inadequate.<sup>5</sup>

In electoral research, the main competitors of the MNL that do not require  $\epsilon_{nj}$  to be iid are the Multinomial Probit Model (MNP, Alvarez and Nagler 1998) and the Mixed Parameters Logit Model (MXL, Train 2009, chap. 6). But although the MNP and the MXL are superior in theory, Whitten and Palmer (1996, 256) argue that major cases of party system change that would expand or contract the choice set are rare and will affect voting behavior so strongly that “inferences drawn from an empirical analysis of an election that occurred prior to party entry or exit would be questionable regardless of the estimation procedure used.”

More generally, Train (2009, 36) claims that correlations among the errors “seem to have less effect when estimating average preferences than when forecasting substitution patterns.” Accordingly, Dow and Endersby (2004) have demonstrated that the IIA assumption rarely poses a problem in multiparty elections. Finally, simulations have shown that the MNL estimator often outperforms the MNP and MXL estimators even if the IIA assumption is seriously violated (Kropko 2010), while the most popular formal tests for the validity of the IIA assumption often disagree and have poor properties even in large samples (Cheng and Long 2007).

Could a violation of the IIA assumption pose a problem for our method? On the one hand, it could be argued that concerns about correlated errors are largely irrelevant in the context of our measure, because we are neither specifying an explanatory model nor making predictions. Rather, we are using the close relationship between the MNL and our measure to obtain estimates of  $A'_i$ s and their standard errors in the same way the binary logit model could be employed to estimate Martin, Traugott, and Kennedy’s  $A$ .

More pragmatically, the MNP and the MXL are not identified for an intercept only setup, as there is no variation within groups from which a covariance structure for  $\epsilon_{nj}$  could be estimated. For the same reason, formal tests such as the Hausman test or the Small-Hsiao test of IIA which work by estimating restricted models that exclude one of the alternatives will always fail to disconfirm<sup>6</sup> the IIA assumption because this restriction does not change the odds in a model containing only intercepts.

<sup>4</sup>The combination relies on an approximation by the delta method that is implemented in the built-in Stata command `nlcom`. For a detailed account, see Section 2 of the supplementary material. The necessary calculations and calls to `nlcom` are carried out by our add-on.

<sup>5</sup>As Train (2009, 42) points out, there are (at least) three conditions under which the logit specification is inadequate: when the effects of attributes vary randomly over decision makers, when repeated choices are affected by correlated unobserved factors, and when substitution across alternatives is not proportional.

<sup>6</sup>The calculation of the Hausman statistic involves pre- and post-multiplication by the vector of differences between the coefficients of the full and the restricted model, which are null. Therefore, the test statistic is always null for the model

On the other hand, however, the validity of the IIA assumption is first and foremost a substantive question that should not be dismissed for technical reasons. We, therefore, urge researchers wishing to apply our methodology to heed McFadden's (1973, 113) early advice that multinomial logit analysis should be restricted to situations where choices are perceived as "distinct . . . by each decision maker." More specifically, differences between polls and electoral results might overstate the degree of bias in a setting where two or more parties are seen as close substitutes for each other by a relevant number of voters (e.g., because they form a pre-electoral alliance or are jointly perceived as "outsider parties") so that voters are essentially indifferent between them. In such cases, researchers should make the IIA assumption more plausible by grouping similar parties together.<sup>7</sup>

### 2.3 Composite Measure of Polling Bias: $B$ and $B_w$

Although a series of  $k$  measures of bias for each and every party is certainly informative, researchers will want to know if the poll as a whole is biased, and whether this bias is statistically significant. Assessing the statistical significance of any differences between sample and population is straightforward using a goodness of fit test employing either Pearson's classic  $\chi^2$  or the likelihood-based  $G^2$  statistic.

Both  $\chi^2$  and  $G^2$  rely on approximations that work well in large samples. For very small samples or situations where the expected values are very low for some categories, exact tests are more appropriate. However, this should rarely be a concern in practice, as modern academic and commercial opinion surveys normally fall into the "large" category. Unfortunately,  $\chi^2$  and  $G^2$  are not suitable for summarizing a poll's total bias. As test statistics, they depend on the sample size and the number of categories and have no clear substantive interpretation.

Simply averaging over the  $A'_i$ s is intuitively appealing but not an ideal solution, because the  $A'_i$ s and their sampling distributions are not independent of each other. More specifically, in a scenario with  $k$  parties, one of the  $k$   $A'_i$ s is redundant, because the projected and real vote shares must sum to unity. To see how the  $A'_i$ s are related, consider  $A'_1$  in a three-party race:

$$A'_1 = \ln\left(\frac{p_1}{1-p_1} \times \frac{1}{O_1}\right) \quad (9)$$

$$= \ln\left(\frac{1-(p_2+p_3)}{(p_2+p_3)} \times \frac{1}{O_1}\right), \quad (10)$$

since  $1-p_1 = p_2+p_3$ .

$p_2$  and  $p_3$  can be reexpressed in terms of  $A'_2$  and  $A'_3$  by reversing the transformation in equation (6):

$$p_i = \frac{\exp(A'_i \times O_i)}{1 + \exp(A'_i \times O_i)}. \quad (11)$$

Substituting equation (11) into (10) yields

$$A'_1 = \ln\left(\frac{1 - \left(\frac{\exp(A'_2 \times O_2)}{1 + \exp(A'_2 \times O_2)} + \frac{\exp(A'_3 \times O_3)}{1 + \exp(A'_3 \times O_3)}\right)}{\frac{\exp(A'_2 \times O_2)}{1 + \exp(A'_2 \times O_2)} + \frac{\exp(A'_3 \times O_3)}{1 + \exp(A'_3 \times O_3)}} \times \frac{1}{O_1}\right). \quad (12)$$

containing only intercepts. The Small-Hsiao test randomly divides the data into subsamples. Again, there will be no systematic difference between the full and the restricted model, but because of the randomization, a proportion of the tests that converges to the chosen significance level  $\alpha$  will reject the null hypothesis.

<sup>7</sup>This is akin to the nested logit model (Train 2009, chap. 4.2).

More generally, each  $A'_i$  can be written as

$$A'_i = \ln \left( \frac{1 - \sum_{j=1}^k \left( \frac{\exp(A'_j \times O_j)}{1 + \exp(A'_j \times O_j)} \right)}{\sum_{j=1}^k \left( \frac{\exp(A'_j \times O_j)}{1 + \exp(A'_j \times O_j)} \right)} \times \frac{1}{O_i} \right) \text{ for } j \neq i. \quad (13)$$

The odds for any party  $i$  can be restated as a function of the odds for all other parties, because the vote shares sum to unity. Since

$$O_i = \frac{v_i}{\sum_{j=1}^k v_j} = \frac{1 - \sum_{j=1}^k v_j}{\sum_{j=1}^k v_j} \text{ for } j \neq i \quad (14)$$

$$\text{and } v_i = \frac{O_i}{1 + O_i}, \quad (15)$$

$$O_i = \frac{1 - \sum_{j=1}^k \frac{O_j}{1 + O_j}}{\sum_{j=1}^k \frac{O_j}{1 + O_j}} \text{ for } j \neq i, \quad (16)$$

which can again be substituted into (13):

$$A'_i = \ln \left( \frac{1 - \sum_{j=1}^k \left( \frac{\exp(A'_j \times O_j)}{1 + \exp(A'_j \times O_j)} \right)}{\sum_{j=1}^k \left( \frac{\exp(A'_j \times O_j)}{1 + \exp(A'_j \times O_j)} \right)} \times \frac{\sum_{j=1}^k \frac{O_j}{1 + O_j}}{1 - \sum_{j=1}^k \frac{O_j}{1 + O_j}} \right) \text{ for } j \neq i. \quad (17)$$

The upshot is that the dependencies within  $\mathbf{p}$  and  $\mathbf{v}$  create a more complicated dependency among the  $A'_i$ s. More precisely,  $\mathbf{p}$  and  $\mathbf{v}$  are *compositions*, because they must sum to the respective total of voters/respondents. This constraint generates negative correlations among the sampling distributions of their constituents, which renders them unsuitable for standard statistical analysis — a problem that was identified by Pearson (1897) in the late nineteenth century but has mostly been ignored in the social sciences (Bacon-Shone 2011).

In  $\mathbf{p}$  and  $\mathbf{v}$ , each element is a linear combination of all the other elements. In the vector of  $A'_i$ s, each element is a *nonlinear* combination of all the other elements, carrying the problem forward. This is obvious for  $k=2$ , where  $A'_2$  is just the negative of  $A'_1$ . Consequently, the average of both values is always zero, regardless of the extent of survey bias. This perfect negative relationship between the  $A'_i$ s is somewhat diluted as  $k$  becomes larger and the distribution of voters/respondents more equal across the categories, but the fact remains that the mean of the  $A'_i$ s gives an overoptimistic impression of the survey's quality, because it is effectively biased toward zero.

Aitchison (1982) has pioneered the use of log-ratio transformations of compositional data that makes multivariate analysis feasible. Building on this seminal contribution, he and others have extended this approach to accommodate a host of complex problems (see Pawlowsky-Glahn and Buccianti 2011). These techniques, however, deal with the transformation of vectors of compositional *data* which are constrained to sum up to a total, whereas our generalization of Martin, Traugott, and Kennedy's  $A$  is a vector of scalar *measures* that are already based on a logged ratio of ratios, subject to a more complex constraint. It is not obvious how compositional data analysis could be applied in this situation.

We therefore propose using the unweighted average of the respective *absolute* values of the  $k$   $A'_i$ s as the aggregate measure of bias,  $B$ :

$$B = \frac{\sum_{i=1}^k |A'_i|}{k}. \quad (18)$$

In a two-party scenario,  $B$  is identical to the absolute value of Martin, Traugott, and Kennedy's original  $A$  and so again retains  $A$ 's useful properties.<sup>8</sup> For  $k$  parties,  $\exp(B)$  is the average factor by which the parties' odds are over- or underestimated. Since the average is not weighted by true party size, including many small parties for which survey bias is small in absolute terms but large in relative terms may result in overly pessimistic  $B$ s.

In forecasting applications, with focus generally on predicting the vote of larger parties, the unweighted  $B$  will be inflated by inaccurate polls for parties which win only very small shares of the vote. To correct this, we offer a weighted version of index,  $B_w$ , which weights contribution to overall poll error by relative share of the total vote. An alternative strategy would be to combine all small parties into a generic "other" category, in the expectation that individual polling errors for these parties will cancel out, and therefore will inflate  $B$  less markedly. For researchers interested in polling performance for small parties, the unweighted  $B$  is evidently more appropriate.

## 2.4 Significance Tests for Multiparty Polling Bias

But  $B$  and  $B_w$  are not without problems. Although the  $A_i$ 's themselves are approximately normally distributed (as shown in Section 2 of the supplementary material), their absolute values follow a folded normal distribution that is skewed to the right and has a positive expectation, as there are by definition no negative values. Their sum (on which the average is based) also has a positive expectation and a nonnormal sampling distribution, although the central limit theorem guarantees that the nonnormality quickly declines for higher values of  $k$ .<sup>9</sup> As a result,  $B$  and  $B_w$  have positive expectations even if the null hypothesis of zero bias holds. This, and their skewed sampling distribution, would render significance tests based on them dubious at best.

We employ a two-pronged strategy to deal with this problem. First, we suggest testing the null hypothesis of zero bias solely on the basis of  $\chi^2$  and  $G^2$ . Because the  $\chi^2$  and likelihood ratio tests are already specified in terms of  $k - 1$  degrees of freedom to account for the compositional nature of  $\mathbf{p}$  and  $\mathbf{v}$ , no further corrections are required.

Second, since  $B$  and  $B_w$  are attractive because they are easily interpretable, we assess the severity of their upward bias through a series of simulated unbiased draws from known populations. For this simulation experiment, we define six scenarios which closely resemble real-world conditions for survey-based research.<sup>10</sup>

Condition (1) represents a balanced three-party system, while conditions (2) and (3) stand for a "two-and-a-half" party system and a three-party system with a single dominant party, respectively. Scenario (4) represents a highly fragmented six-party system, while scenario (5) stands for a situation with one dominant and many minor parties. Finally, scenario (6) depicts a system with two major, two minor, and two small parties. We refrained from simulating random sampling under systems with even more (relevant) parties, because these would most likely form electoral alliances or be grouped together by the analyst.

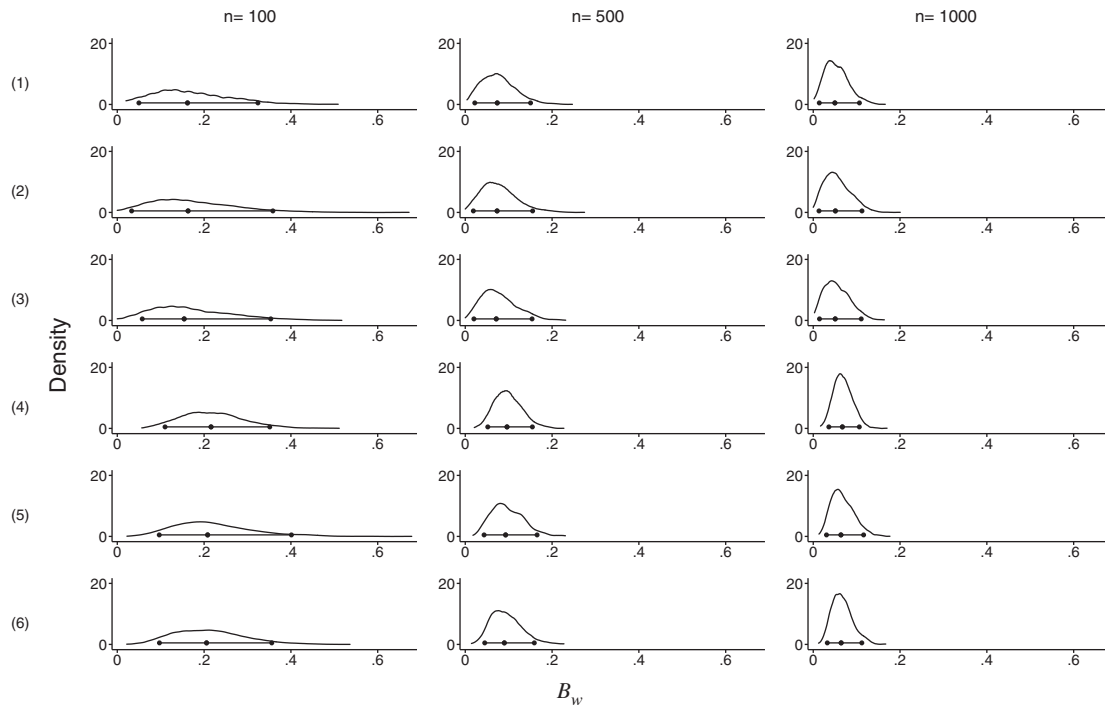
Under each of these conditions, we simulated polling using three sample sizes:  $n = 1000$ , which probably comes closest to what could be considered the standard size of an opinion survey;  $n = 500$ , which might be an adequate size for a representative pilot study; and  $n = 100$ , which could be used in experimental settings or when testing new instruments in the classroom. Each of the eighteen experiments was replicated a thousand times.

Figure 1 shows kernel density estimates of the distribution of  $B_w$  under these conditions, with the thin line under the distribution indicating a centered 90% interval and the small dot on that line marking the median of the distribution. Obviously, for all conditions and sample sizes, the distributions are skewed to the right, although they become more symmetric as the sample size increases.

<sup>8</sup>In a two-party race, bias in favor of the first party necessarily results in bias against the second party and vice versa. The log-transformation ensures that the arbitrary decision which party forms the base for the calculation of the odds affects only the sign of  $A$ , not its magnitude.

<sup>9</sup>Moreover, taking absolute values does not (completely) remove the collinearity but rather changes the signs of the correlations, thereby increasing the variance of the sum's sampling distribution.

<sup>10</sup>These conditions are (1)  $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ ; (2)  $\frac{2}{5}, \frac{2}{5}, \frac{1}{5}$ ; (3)  $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$ ; (4)  $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$ ; (5)  $\frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}$ ; (6)  $\frac{3}{10}, \frac{3}{10}, \frac{1}{20}, \frac{1}{20}, \frac{3}{20}, \frac{3}{20}$ .



**Fig. 1** Simulated sampling distribution of  $B_w$  when null hypothesis of no bias holds.

More importantly, however, the median value of  $B_w$  is reasonably close to zero even if  $n$  is just 500. For  $n = 1000$ , even the 95th percentile is in the range of 0.1, meaning that it is quite unlikely to observe a higher value of this statistic when the null hypothesis of no bias holds.<sup>11</sup> Although its nonzero expectation is clearly a drawback that renders the statistic unfit for significance tests, the results show that the statistic's actual upward bias is moderate in large samples, making  $B_w$  an easily interpretable (if slightly conservative) measure of overall poll accuracy.<sup>12</sup>

### 3 An Application: Polling Bias in French Pre-Election Polls

We turn now to a simple empirical application of the  $B$  and  $B_w$  indices. The 2012 French presidential elections on April 22 saw ten candidates compete in the first round of the election, the top two candidates progressing to the runoff. Table 1 lists the candidate names and eventual first-round scores.

The database of surveys in France in the pre-election period provides a total of 102 datapoints from July 1, 2011. We include only those surveys taking place after March 22, 2012 (one calendar month before the election). This date is also one week after the final cutoff date for candidate registration, after which we expect voter preferences to stabilize relatively quickly. This consequently allows a reasonable number of surveys (32) across eight different polling companies, as well as minimizing likely changes in candidate preference.

#### 3.1 Across-Time Polling Bias in Pre-Election Polls by Organization

Table 1 also indicates the four different coding strategies we use to test the two indices. The Stata add-on we use to calculate the  $B$  measure potentially allows us to include up to twelve individual

<sup>11</sup>A value of 0.1 for  $B_w$  would imply that on average, the odds of a vote for any given party would be over- or underestimated by about 10%, as  $\exp(0.1) \approx 1.105$  and  $\exp(-0.1) \approx 0.905$ .

<sup>12</sup>Finally, one might also wish to consider the test's power to detect overall bias. We carried out a series of simulations which show that the test has sufficient power to pick up substantively relevant bias under a wide range of conditions. Detailed results are presented in Section 3 in the supplementary material.



**Table 1** Percentage of valid vote in first round of 2012 French presidential election

Candidate	% (vote)	Individually coded			
		2	3	5	8
François Hollande	28.63	•	•	•	•
Nicolas Sarkozy	27.18	•	•	•	•
Marine Le Pen	17.90		•	•	•
Jean-Luc Mélenchon	11.10			•	•
François Bayrou	9.13			•	•
Eva Joly	2.31				•
Nicola Dupont-Aignan	1.79				•
Philippe Poutou	1.15				•
Nathalie Arthaud	0.56				
Jacques Cheminade	0.25				

values. Realistically, however, this produces a host of  $A'_i$  scores, which for at least the four lowest-ranked candidates will have little analytical value, unless polls demonstrated very large bias. Two minor Radical Left candidates, Philippe Poutou and Nathalie Arthaud, as well as Jacques Cheminade, a maverick candidate and head of the LaRouche movement in France, and Nicolas Dupont-Aignan, a dissident right-winger, only managed 3.75% of the vote between them, and so would normally be classified “other.” There is also an evident differential between the fifth and sixth candidates, and indeed between the second and third. Of greater interest, then, would be what happens to the  $B$  index when different coding choices are made. We therefore include four different scenarios, with two-, three-, five-, and eight-candidate codings. For the purposes of analysis, we treat each of the four different codings as an observation, giving a total of 128 observations across the thirty-two surveys.

Using the Pearson  $\chi^2$  and likelihood ratio  $G^2$  scores, a total of 68, or just over half of the observations, turn out to be biased at the 95% confidence level.<sup>13</sup> Table 2 reports the  $B$  and  $B_w$  values for these. The two-candidate column is omitted, as none of the surveys showed significant bias using this coding. Variation from the eventual outcome among the losing candidates (those missing the second round runoff) clearly nets out. Equally, a perverse effect of coding these eight candidates in a single category turns them into an artifactual largest “single” candidate, thereby weighting down the individual error for the two real candidates in  $B_w$ .

The three-candidate coding shows much greater evidence of bias—entirely in line with expectations, given that the extreme right candidate, Marine Le Pen, displayed much more unstable polling estimates throughout the entire campaign. The five-candidate coding, including all candidates with scores above 5%, sees the vast majority of surveys (26 of 32) exhibit significant bias. Results for the even more fine-grained eight-candidate coding are very similar, with 24 of 32 surveys deviating significantly from the election’s result.<sup>14</sup>

Statistical significance is, however, not very interesting in itself. More importantly, bias is relatively small in absolute terms. The average values of  $B_w$  for the two-, three-, five-, and eight-candidate codings are 0.055, 0.104, 0.122, and 0.132, respectively. Put differently, the French pollsters did a reasonable job predicting the outcome of the first round.

Individual outliers are noticeable, though. For example, the CSA polls in late March and early April show the highest bias, through an overestimation of Sarkozy’s (30%) and Mélenchon’s (15%) scores and underestimating Le Pen’s (13%). However, this organization’s scores come into line in

<sup>13</sup>A survey was classified as biased if either of the measures indicated statistically significant bias.

<sup>14</sup>The very slight decrease in the number of significantly biased surveys is due to two LH2 polls (April 11 and April 18). Their borderline significant bias is increased just a little further by coding eight candidates separately, but this change is offset by the additional degrees of freedom.

**Table 2** Statistically significant bias in French pre-election polls

<i>Date</i>	<i>Pollster</i>	$B_w(3)$	$B_w(5)$	$B_w(8)$
March 22, 2012	BVA	0.131	0.182	0.195
March 24, 2012	IPSOS		0.108	0.110
March 25, 2012	IFOP	0.121	0.146	0.163
March 26, 2012	Harris interactive		0.110	0.129
March 27, 2012	TNS Sofres		0.124	0.129
March 27, 2012	OpinionWay		0.087	0.105
March 27, 2012	CSA	0.148	0.187	0.196
March 31, 2012	BVA	0.098	0.110	0.115
March 31, 2012	LH2	0.120	0.168	0.173
March 31, 2012	IPSOS	0.133	0.160	0.165
April 02, 2012	Harris interactive	0.123	0.138	0.157
April 02, 2012	CSA	0.133	0.200	0.210
April 04, 2012	OpinionWay	0.122	0.146	0.147
April 07, 2012	IPSOS		0.137	0.137
April 11, 2012	LH2	0.110	0.102	
April 11, 2012	CSA	0.150	0.176	0.182
April 12, 2012	BVA		0.124	0.138
April 12, 2012	TNS Sofres		0.110	0.120
April 14, 2012	IPSOS	0.112	0.101	0.119
April 15, 2012	IFOP	0.084	0.087	0.103
April 17, 2012	CSA	0.109	0.122	0.122
April 17, 2012	BVA	0.105	0.143	0.151
April 18, 2012	LH2	0.123	0.112	
April 19, 2012	CSA	0.126	0.114	0.125
April 19, 2012	BVA	0.123	0.126	0.151
April 20, 2012	IFOP		0.076	0.081

later polls, even if still biased, and indeed scores for CSA polls before our start-date were better (see Fig. 1 in the supplementary material).

### 3.2 Using $B$ and $B_w$ to Model Explanations of Polling Bias

Such outliers raise an important issue with individual polling organizations and sources of bias. Much of the polling literature discusses the role of “house effects” in causing consistent bias in one or more parties’ scores (Jackman 2005; Fisher et al. 2011), although a house effect could equally focus on average stability of estimates over time. There are, however, at least three other potential causes of bias which need to be controlled for to allow a robust estimation of their relative strengths and of house effects. First, the number of days before the election should capture any remaining fluctuations in public opinion in the campaign, which may cause changes in intended vote or last-minute decisions. Averaging across surveys should allow an estimate of the amount of convergence for this election. Second, the sample size affects significance testing but as the simulations have shown,  $B_w$  is also biased away from zero—the larger the sample size, the smaller the bias, other things being equal. Finally, even with stable sample size, we have seen that candidate coding choices affect the significance and level of bias, with smaller numbers of individual candidates being less prone to bias. For a multi-party system, then, the use of  $B$  or  $B_w$  as the outcome allows us to model these different sources of bias.

In Table 3, we apply this to the French presidential data. Each poll for which we have full data constitutes an observation. The dependent variable is simply the  $B_w$  score for each poll. The house effect is coded as a series of dummy variables for polling organization, with polls by the Opinionway organization as the reference. The sample size, coding of number of candidates, and length of time to the election could be included as simple covariates. However, while the direction

**Table 3** Polynomial regression of bias on number of candidates, sample size, time, and polling company

	$B_w$
Candidates <sup>-2</sup> – 0.05	–0.325*** (0.0209)
N-1148	–0.0000217** (0.00000715)
(days/10) <sup>2</sup> – 2.02	0.0201*** (0.00458)
(days/10) <sup>3</sup> – 2.87	–0.00596*** (0.00159)
BVA	0.0375*** (0.00858)
CSA	0.0504*** (0.00729)
Harris	0.00317 (0.00766)
IFOP	0.0247* (0.0104)
IPSOS	0.0102 (0.00730)
LH2	0.0170* (0.00828)
TNS Sofres	0.00522 (0.00832)
Constant	0.107*** (0.00575)
$R^2$	0.755
$N$	128

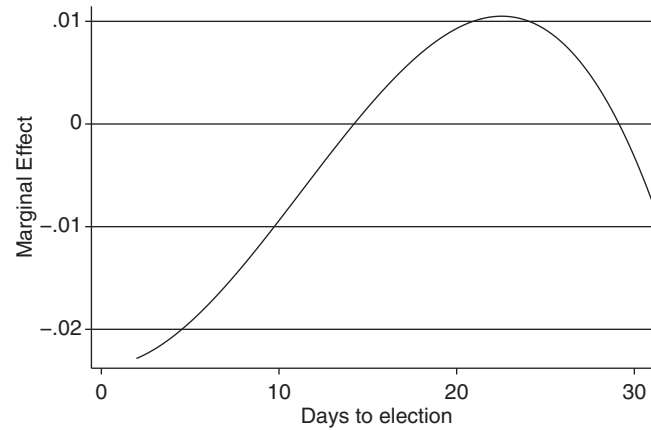
Note. Standard errors in parentheses. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$

of their expected effects is clear, their functional form is not necessarily linear. To account for this, we run a polynomial regression using Stata's multivariable fractional polynomial command prefix. Fractional polynomials provide a flexible yet parsimonious framework to accommodate different types of nonlinearity (Royston and Sauerbrei 2008). We let the Royston-Sauerbrei algorithm choose the optimal transformation for the three controls.<sup>15</sup> Because of the stacked data structure that includes four observations for each unit of analysis, robust standard errors are used to account for clustering.

Table 3 gives the model estimates. The three hypothesized causes of bias potentially confounding house effects all reach significance in the expected direction. Length of time to the election has a distinctly nonlinear effect that is represented by a quadratic and a cubic term (Fig. 2), with greater accuracy in the final few days of the campaign. Somewhat surprisingly, inaccuracy peaks three weeks before election day, with bias levels four and two weeks before the end of the campaign being roughly comparable. This might be due to some campaign effects on political opinion, which may later have returned to equilibrium. The peculiar shape of the curve should not, however, be overinterpreted, as the marginal effect of time is generally rather small.

Similarly, increasing sample size only very moderately reduces bias. As Table 3 shows, for every 1000 additional respondents, the  $B_w$  score is expected to reduce by around 0.0217. One should still

<sup>15</sup>Following Royston and Sauerbrei, the transformed variables are centered to reduce collinearity. The structure of the model itself remains linear-additive.



**Fig. 2** The marginal effect of number of days to the election on polling bias (fractional polynomial effect).

bear in mind that all sample sizes under study are relatively large.<sup>16</sup> For smaller samples, the effect of sample size would most likely be more pronounced and nonlinear.

Finally, Fig. 3 demonstrates the effect of model structure, with the fit across the different candidate codings. In line with the individual  $B_w$  scores, there is a more pronounced rise in inaccuracy as the candidate coding moves from two to three, with a less steep increase as more and more minor candidates are coded individually. This is also undoubtedly due to the downward weighting of minor candidates in  $B_w$ 's calculation. Of course, this particular curve is a function of the party system distribution—the relative sizes of each candidate's vote share. For different shares, the exact shape of the function will differ. However, it does demonstrate that those using the  $B_w$  index should be aware of methodological artifacts in their findings of more or less accuracy. To make such decisions intelligently, a common benchmark, such as a fractionalization or effective number index (e.g., Laakso and Taagepera 1979) or a proportion of vote threshold, might be used to indicate the number of relevant competitors.<sup>17</sup>

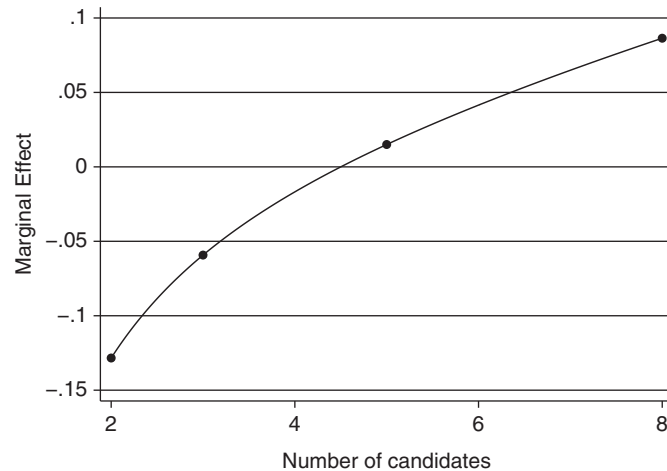
#### 4 Conclusion

In a period when polling accuracy is coming under increased scrutiny, both in terms of polls' ability to forecast elections (even if individual polls are not forecasts per se) and the underlying biases inherent in their method which cause inaccuracy, the lack of a single index able to characterize polling accuracy for multi-party systems is problematic. Though not the only reason for less study of multi-party systems in the polling literature, the complexities of looking at multiple, related indicators to track polling accuracy has certainly contributed to putting these polls beyond the scope of more formalized approaches such as those found in the literature on UK and US elections. Furthermore, in the wake of a US election where the robust use of polling data by researchers like Nate Silver, Drew Linzer, and Simon Jackman publicly demonstrated the value of polling estimates informing electoral commentary, the lack of a simple unified index which would allow even retrospective review of overall polling performance in the majority of the world's democracies is a significant impediment to developing an understanding of such opinion measures in political behavior.

Not only has the lack of a measure reduced the capacity of analysts to understand patterns in polling data, it has also meant there is no yardstick by which to measure the performance of polling institutions in elections. In European democracies where concerns over polls influencing voters via a feedback loop are such that some commentators worry that they are degrading democracy (e.g.,

<sup>16</sup>N varies from 876 to 2555.

<sup>17</sup>In the French example, the effective number of candidates was 4.8 (Evans 2012, 124). That would suggest coding five candidates plus "others."



**Fig. 3** The marginal effect of the independently coded number of candidates on polling bias (fractional polynomial effect).

Italy's purported *sondocrazia*), the ability to identify the location and more fundamentally the degree of inaccuracy is vital.

The  $B$  index of course cannot be used for *ex ante* forecasts of elections, in keeping with all measures of predictive accuracy—an observed outcome is required. Moreover, pre-election polls are not forecasts of actual election outcomes, but rather snapshots of public opinion at points in time prior to the election. As we have demonstrated, our measure allows researchers to track the evolution of polling across time, and to use this measure as a dependent variable. Indeed, this potential use was noted by Martin, Traugott, and Kennedy for their two-party index (2005, 352).

Inevitably, the underlying mathematics of the multi-party measure are more complex than the two-party measure, but the index itself, as well as the individual  $A'_i$  scores, are easily estimated and interpreted. As with any index, choices that the researcher makes in terms of number of individual candidates/parties to retain and whether  $B$  or  $B_w$  is more appropriate may influence their substantive findings. Replication by other researchers of these analyses, using different candidate or party groupings, will reveal the effect of such choices on conclusions about accuracy.

Historically, measures of polling accuracy such as the Mosteller measures have followed elections where the inaccuracy of pre-election polls was thought to be problematic. There have been numerous instances in recent multi-party elections in Europe and elsewhere of polls providing misleading trends (Schaffer and Schneider 2005; Callegaro and Gasperoni 2008; Durand 2008). We hope that  $B$ 's use could provide an incentive for all polling organizations to supply the requisite information for researchers to calculate the index, to allow transparency in assessing polls' relative performance across a campaign.

## References

- Agresti, Alan. 2002. *Categorical data analysis*. 2nd ed. Hoboken, NJ: John Wiley.
- Aitchison, John. 1982. The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B* 44(2):139–77.
- Alvarez, R. Michael, and Jonathan Nagler. 1998. When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science* 42:55–96.
- Arzheimer, Kai, and Jocelyn Evans. 2013. Replication data for: A new multinomial accuracy measure for polling bias. *IQSS Dataverse Network/Distributor V1* [Version]. <http://hdl.handle.net/1902.1/21603> (accessed August 6, 2013).
- Bacon-Shone, John. 2011. A short history of compositional data analysis. In *Compositional data analysis: Theory and applications*, eds. Vera Pawlowsky-Glahn and Antonella Buccianti, 1–11. Chichester, UK: Wiley.
- Callegaro, Mario, and Giancarlo Gasperoni. 2008. Accuracy of pre-election polls for the 2006 Italian parliamentary election: Too close to call. *International Journal of Public Opinion Research* 20(2):148–70.
- Cheng, Simon, and J. Scott Long. 2007. Testing for IIA in the multinomial logit model. *Sociological Methods & Research* 35(4):583–600.

- Dow, Jay K., and James W. Endersby. 2004. Multinomial probit and multinomial logit: A comparison of choice models for voting research. *Electoral Studies* 23:107–22.
- Durand, Claire. 2008. The polls of the 2007 French presidential campaign: Were lessons learned from the 2002 catastrophe? *International Journal of Public Opinion Research* 20(3):275–98.
- Evans, Jocelyn. 2012. The sound foundations of a socialist victory. *Renewal* 20(2–3):123–28.
- Fisher, Stephen D., Robert Ford, Will Jennings, Mark Pickup, and Christopher Wlezien. 2011. From polls to votes to seats: Forecasting the 2010 British general election. *Electoral Studies* 30(2):250–57.
- Jackman, Simon. 2005. Pooling the polls over an election campaign. *Australian Journal of Political Science* 40(4):499–517.
- Kropko, Jonathan. 2010. A comparison of three discrete choice estimators (unpublished dissertation chapter). University of North Carolina-Chapel Hill. <http://www.unc.edu/~kropko/paper1.pdf> (accessed August 6, 2013).
- Laakso, Markku, and Rein Taagepera. 1979. “Effective” number of parties: A measure with application to West Europe. *Comparative Political Studies* 12(1):3–27.
- Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Luce, R. Duncan. 1959. *Individual choice behavior. A theoretical analysis*. New York: John Wiley & Sons.
- Martin, Elizabeth A., Michael W. Traugott, and Courtney Kennedy. 2005. A review and proposal for a new measure of poll accuracy. *Public Opinion Quarterly* 69(3):342–69.
- McFadden, Daniel. 1973. Conditional logit analysis of qualitative choice behaviour. In *Frontiers of Econometrics*, ed. Paul Zarembka, 105–42. New York: Academic Press.
- McLean, Iain. 1995. Independence of irrelevant alternatives before Arrow. *Mathematical Social Sciences* 30(2):107–26.
- Mosteller, Frederick, Herbert Hyman, Philip McCarthy, Eli Marks, and David Truman. 1949. *The pre-election polls of 1948: Report to the committee on analysis of pre-election polls and forecasts*. New York: Social Science Research Council.
- Pawlowsky-Glahn, Vera, and Antonella Bucciante, eds. 2011. *Compositional data analysis: Theory and applications*. Chichester, UK: Wiley.
- Pearson, Karl. 1897. Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proceedings of the Royal Society of London* 60:489–502.
- Ray, Paramesh. 1973. Independence of irrelevant alternatives. *Econometrica* 41(5):987–91.
- Royston, Patrick, and Willi Sauerbrei. 2008. *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester, UK: Wiley.
- Schaffer, Lena-Maria, and Gerald Schneider. 2005. Die Prognosegte von Wahlbörsen und Meinungsumfragen zur Bundestagswahl 2005. *Politische Vierteljahresschrift* 46(4):674–81.
- Train, Kenneth. 2009. *Discrete choice methods with simulation*. 2nd ed. Cambridge, UK: Cambridge University Press.
- Whitten, Guy D., and Harvey D. Palmer. 1996. Heightening comparativists’ concern for model choice: Voting behavior in Great Britain and the Netherlands. *American Journal of Political Science* 40(1):231–60.