A New Multinomial Accuracy Measure for Polling Bias

Kai Arzheimer Jocelyn Evans

Abstract

In this paper, we propose a polling accuracy measure for multi-party elections based on a generalisation 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 organisations. 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.

This is the authors' version. The final version will appear in Political Analysis 2013. Go to `www.kai-arzheimer.com/a-new-multinomial-accuracy-measure-for-polling-bias` for more information.

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. Illustrating their approach with the competition between the two 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 favouring 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,[1] 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.

[1] Or more generally to dichotomous variables with a known distribution in the population, such as gender or registration as a voter.

2.2 Generalisation for the Multi-Party Case

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

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

Note that p_1, p_2, v_1, v_2 could be written as a 2×2 table: $\left(\begin{smallmatrix} p_1 & p_2 \\ v_1 & v_2 \end{smallmatrix}\right)$. A is then identical to the log of the familiar odds ratio, 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 odds ratios (Agresti, 2002, 56). This suggests calculating a series of logged odd ratios 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 odds ratio 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

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

which results in a negative log odds ratio although support for party a is measured without bias. In short, a measure based on the (log) odds ratio would be biased proportionate to the reference party's own polling bias. Secondly, it follows that, because the (log) odds ratios 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.

[2] We can treat (self-declared) non-voters as an additional party.

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

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

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 generalise A for use in a multi-party competition:

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

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

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

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 favour 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.[3]

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 realisation 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 multi-party system, and consequently no software package available to calculate any such estimators. It

[3]Moreover, like in the case of the original A, calculating $\exp(A'_i)$ recovers the odds ratio.

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 parametrised 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

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

where β_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 parametrisation into Equation (5), we can calculate A_2' as

$$A_2' = \ln\left(\frac{\frac{\exp(\beta_2)}{\sum_{j=1}^{k} \exp(\beta_j)}}{1 - \frac{\exp(\beta_2)}{\sum_{j=1}^{k} \exp(\beta_j)}} \times \frac{1 - v_2}{v_2} \right) \tag{8}$$

While re-defining A_i' as a non-linear function of $k-1$ parameters might appear complicated, estimating the βs 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'.[4]

Inevitably, use of the MNL may raise concerns over assumptions of independence from irrelevant alternatives (IIA). The notion of independence from irrelevant alternatives has a long and convoluted history in game theory and (social) choice theory (Ray, 1973) that harks back to the 18th 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

[4]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 B. Again, the necessary calculations and calls to `nlcom` are carried out by our add-on.

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 ϵ_{nj} be some additional unobserved utility component (Train, 2009, 14-15). Finally, assume that while 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 ϵ_{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.[5]

In electoral research, the main competitors of the MNL that do not require ϵ_{nj} to be iid are the Multinomial Probit Model (MNP, Alvarez and Nagler 1998) and the Mixed Parameters Logit Model (MXL, Train 2009, ch. 6). But while 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 behaviour 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 amongst 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 multi-party 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 methodology? 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

[5] 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.

relationship between the MNL and our measure to obtain estimates of A_i', B, 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 ϵ_{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[6] the IIA assumption because this restriction does not change the odds in a model containing only intercepts.

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.[7]

2.3 Composite Measure of Polling Bias: B and B_w

While 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 χ^2 or the likelihood-based G^2 statistic.

Both χ^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

[6]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 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 randomisation, a proportion of the tests that converges to the chosen significance level α will reject the null hypothesis.

[7]This is akin to the nested logit model (Train, 2009, ch. 4.2).

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, χ^2 and G^2 are not suitable for summarising 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 is intuitively appealing but not an ideal solution, because the A'_is and their sampling distributions are not independent of each other. More specifically, in a scenario with k parties, one of the k A'_is is redundant, because the projected and real vote shares must sum to unity. To see how the A'_Is are related, consider A'_1 in a three-party race:

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

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

since $1 - p_1 = p_2 + p_3$.

p_2 and p_3 can be re-expressed in terms of A'_2 and A'_3 by reversing the transformation in Equation (6):

$$p_i = \frac{\exp\left(A'_i \times O_i\right)}{1 + \exp\left(A'_i \times O_i\right)} \tag{11}$$

Substituting Equation (11) into (10) yields

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

More generally, each A'_i can be written as

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

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

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

$$\text{and } v_i = \frac{O_i}{1 + O_i}, \tag{15}$$

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

which can again be substituted into (13):

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

The upshot is that the dependencies within **p** and **v** create a more complicated dependency amongst the A'_is. More precisely, **p** and **v** are *compositions*, because they must sum to the respective total of voters/respondents. This constraint generates negative correlations amongst 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 19th century but has mostly been ignored in the social sciences (Bacon-Shone, 2011).

In **p** and **v**, each element is a linear combination of all the other elements. In the vector of A'_is, each element is a *non*linear 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'_is 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'_is gives an over-optimistic impression of the survey's quality, because it is effectively biased towards 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 generalisation 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'_is as the aggregate measure of bias, B:

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

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.[8] For k parties, $\exp(B)$ is the average factor by which the parties' odds are over- or under-estimated. 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 Bs.

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 Multi-Party Polling Bias

But B and B_w are not without problems. While the A'_is themselves are approximately normally distributed, (as shown in Section B), 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 non-normal sampling distribution, although the central limit theorem guarantees that the non-normality quickly de-

[8] In a two-party race, bias in favour 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.

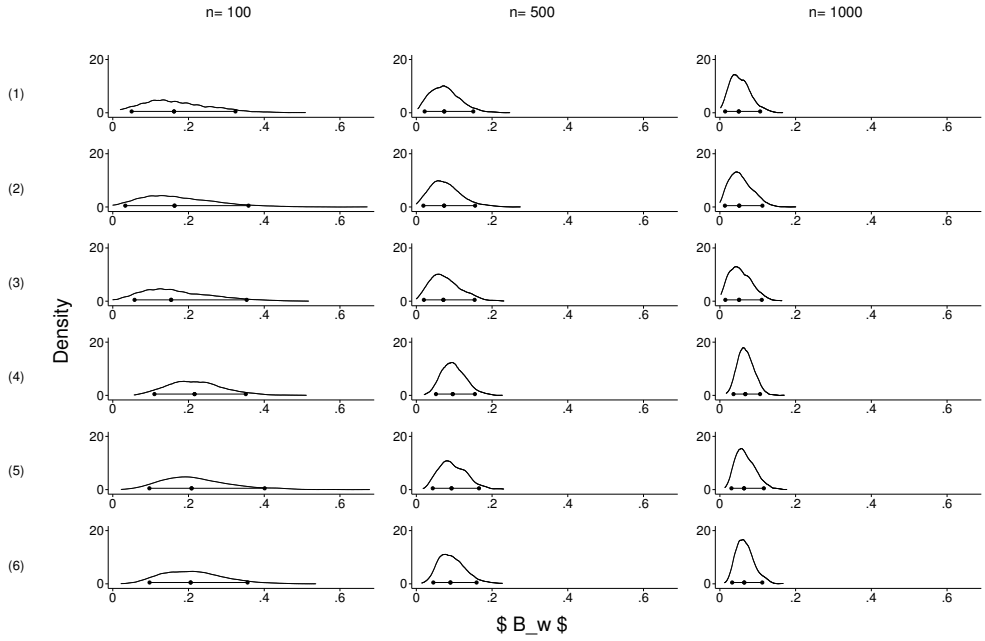

Figure 1: Simulated sampling distribution of B_w when null hypothesis of no bias holds

clines for higher values of k.[9] 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 χ^2 and G^2. Because the χ^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.[10]

[9]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.

[10]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) $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$;

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 18 experiments was replicated a thousand times.

Figure 1 shows kernel density estimates of the distribution of B_w under these conditions, with the thin black line under the distribution indicating a centred 90 per cent interval and the small black 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. 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.[11] While its non-zero 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.[12]

(5) $\frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}$; (6) $\frac{3}{10}, \frac{3}{10}, \frac{1}{20}, \frac{1}{20}, \frac{3}{20}, \frac{3}{20}$.

[11] 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 ten per cent, as $\exp(0.1) \approx 1.105$ and $\exp(-0.1) \approx 0.905$.

[12] 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 C.

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				

Table 1: Percentage of valid vote in first round of 2012 French Presidential election

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 22 April saw 10 candidates compete in the first round of the election, the top two candidates progressing to the run-off. 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 1 July 2011. We include only those surveys taking place after 22 March 2012 (one calendar month before the election). This date is also one week after the final cut-off date for candidate registration, after which we expect voter preferences to stabilise relatively quickly. This consequently allows a reasonable number of surveys (32) across eight different polling companies, as well as minimising likely changes in candidate preference.

3.1 Across-time Polling Bias in Pre-Election Polls by Organisation

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 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 32 surveys.

Using the Pearson χ^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.[13] 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 amongst the losing candidates (those missing the second round run-off) clearly nets out. Equally, a perverse effect of coding these eight candidates in a single category turns them into an artefactual 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.[14]

[13] A survey was classified as biased if either of the measures indicated statistically significant bias.

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

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

Table 2: Statistically significant bias in french pre-election polls

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 (30%) and Mélenchon's (15%) scores and underestimating Le Pen's (13%). However, this organisation's scores come into line in later polls, even if still biased, and indeed scores for CSA polls before our start-date were better (see Figure 4).

3.2 Using B and B_w to Model Explanations of Polling Bias

Such outliers raise an important issue with individual polling organisations 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. Secondly, 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. Lastly, 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 organisation, with polls by the Opinionway organisation 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, whilst the direction 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 non-linearity (Royston and Sauerbrei, 2008). We let the Royston-Sauerbrei algorithm choose the optimal transformation for the three controls.[15] Because of the stacked data structure that includes four observations for each unit of analysis for, robust standard errors are used to account for clustering.

Table 3 gives the model estimates. The three hypothesised causes of bias potentially confounding house effects all reach significance in the expected direction. Length of time to the election has a distinctly non-linear effect that is represented by a quadratic and a cubic term (Figure 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 over-interpreted, 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 bear in mind that all sample sizes under study are relatively large.[16] For smaller samples, the effect of sample size would most likely be more pronounced and non-linear.

Lastly, Figure 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 artefacts in their findings of more or less accuracy. To make such decisions intelligently, a common benchmark, such as a fractionalisation or effective number index (e.g. Laakso and Taagepera, 1979) or a proportion of vote threshold might

[15] Following Royston and Sauerbrei, the transformed variables are centred to reduce collinearity. The structure of the model itself remains linear-additive.

[16] N varies from 876 to 2555.

	B_w
candidates^{-2} − 0.05	-0.325***
	(0.0209)
N-1148	-0.0000217**
	(0.00000715)
(days/10)^2 − 2.02	0.0201***
	(0.00458)
(days/10)^3 − 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

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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

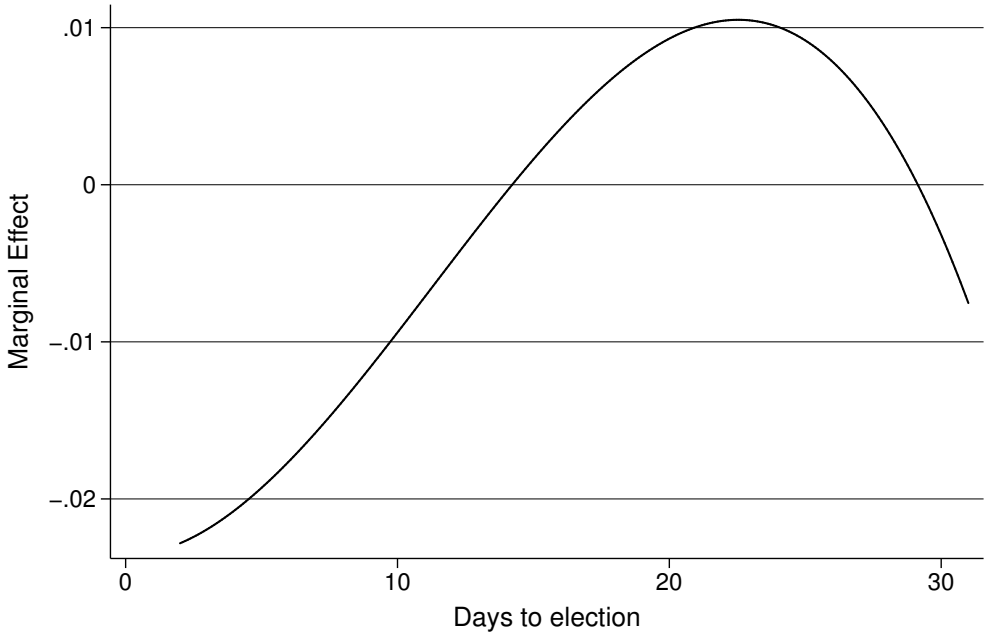

Figure 2: The marginal effect of number of days to the election on polling bias (fractional polynomial effect)

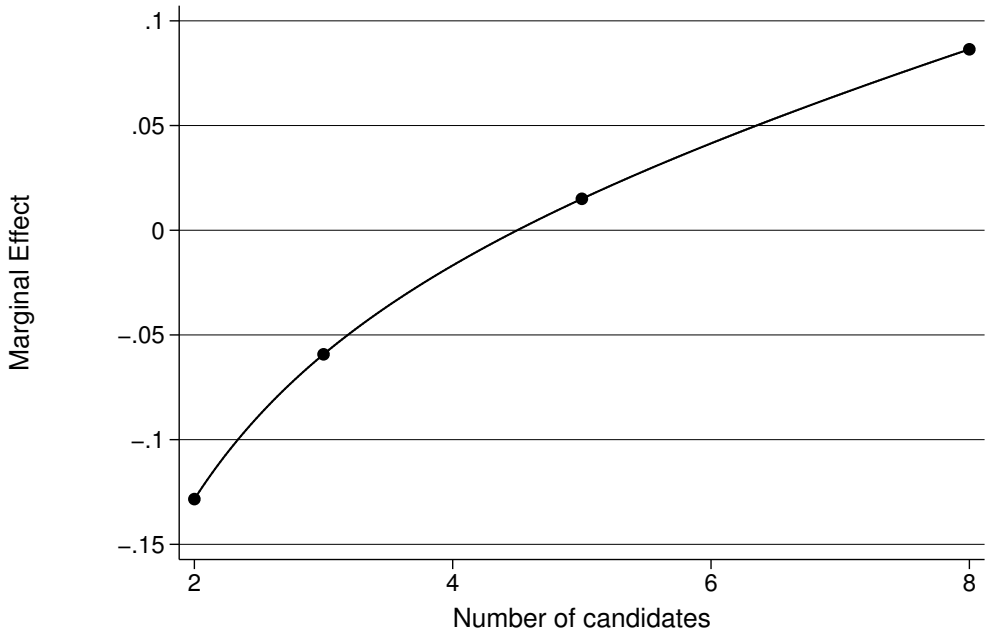

Figure 3: The marginal effect of the independently coded number of candidates on polling bias (fractional polynomial effect)

be used to indicate the number of relevant competitors.[17]

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 characterise polling accuracy for multiparty systems is problematic. Though not the only reason for less study of multiparty 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 formalised approaches such as those found

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

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 behaviour.

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 (Italy's purported *sondocrazia*, for example), 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. Morever, 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's 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 organisations to supply the requisite information for researchers to calculate the index, to allow transparency in assessing polls' relative performance across a campaign.

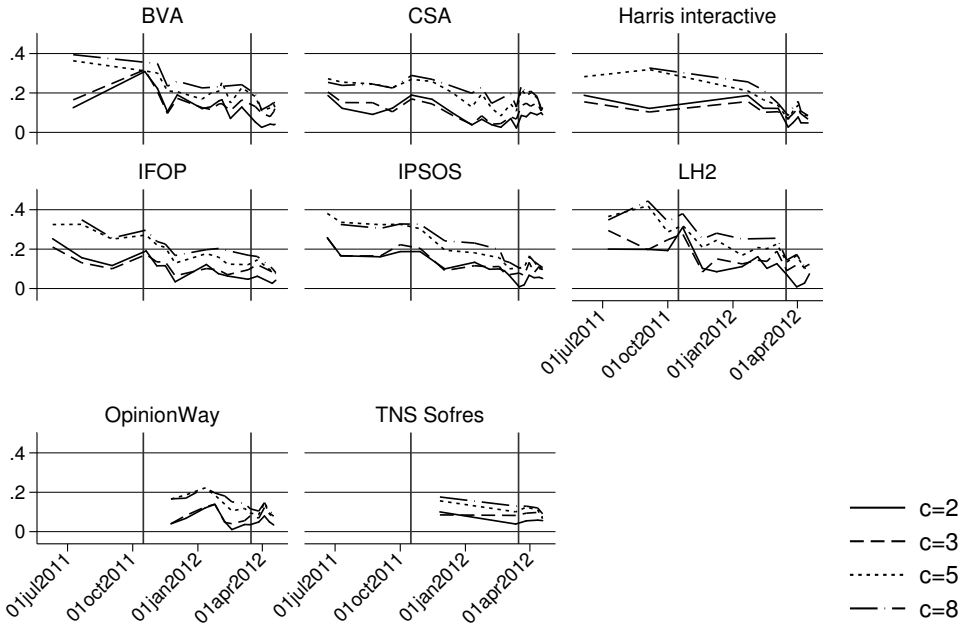

Figure 4: Development of B_w over time for eight French pollsters

A Additional Figures

B Derivation of Standard Errors

In Equation (7) and (8) we demonstrate how A_i' is related to the parameters $\beta_1, \beta_2, \cdots \beta_{k-1}$ of the familiar MNL. Since these calculations are based on survey data, the βs and A_i's should be treated as random vectors, whose variability is of interest.

Because the underlying MNL does not involve any explanatory variables, $\boldsymbol{\beta}$ could in principle be obtained by substituting the observed probabilities into (7) and solving for the $k-1$ unique βs (one β is always set to 0 as an identifying restriction). However, the most convenient way to obtain $\boldsymbol{\beta}$ and the associated variance-covariance matrix $\mathbf{V}$ is to rely on off-the-shelf maximum likelihood procedures for estimating the parameters of the MNL, which are available in all major statistical packages.

For large samples, $\beta_2, \cdots \beta_k$ are distributed multivariate normal. Their

variance-covariances matrix $\mathbf{V}$ can be estimated by the inverse of the information matrix (Agresti, 2002, 193), which is the negative expectation of the matrix of second-order partial derivatives of the likelihood function with respect to the parameters (Hesse Matrix). In that sense, maximum likelihood estimation generates the complete variance-covariance matrix as a byproduct.

Since $A'_1, A'_2, \cdots A'_k$ are defined as non-linear combinations of the βs (see Equation (8)), the derivation of their standard errors is by no means straightforward, although they should be approximately normal as well. To see why this is the case, consider (without loss of generality) the calculation of A'_i for a three party system. Taking the first party as the reference category (for which β_1 is set to 0 so that $\exp(\beta_1) = 1$), A'_1, A'_2 and A'_3 are defined as

$$A'_1 = \ln\left(\frac{\frac{1}{1+\exp(\beta_2)+\exp(\beta_3)}}{1-\frac{1}{1+\exp(\beta_2)+\exp(\beta_3)}} \times \frac{1-v_1}{v_1}\right) \tag{19}$$

$$A'_2 = \ln\left(\frac{\frac{\exp(\beta_2)}{1+\exp(\beta_2)+\exp(\beta_3)}}{1-\frac{\exp(\beta_2)}{1+\exp(\beta_2)+\exp(\beta_3)}} \times \frac{1-v_2}{v_2}\right) \tag{20}$$

$$A'_3 = \ln\left(\frac{\frac{\exp(\beta_3)}{1+\exp(\beta_2)+\exp(\beta_3)}}{1-\frac{\exp(\beta_3)}{1+\exp(\beta_2)+\exp(\beta_3)}} \times \frac{1-v_3}{v_3}\right). \tag{21}$$

Since both β_2 and β_3 are normally distributed, $\exp(\beta_2)$ and $\exp(\beta_3)$ have log-normal distributions, whose location and shape are governed by the respective mean and variance of the underlying normal distribution. For the distribution of the sum of log-normally distributed random variables (i.e. $\exp(\beta_2)+\exp(\beta_3)$ in this case), there is no closed form expression, though it can often be approximated as yet another log-normal distribution. If this approximation holds, the inner fraction also has a log-normal distribution, because the ratio of two log-normal distributions as well as the inverse of a single log-normal distribution are also distributed log-normal.

Things are further complicated by the addition of 1 in the denominator of the inner fractions, the covariation between β_1 and β_2, the difference in the denominator of the outer fraction, and the multiplication by the odds of the vote. In short, while there is reason to believe that the expression within the parentheses is distributed approximately log-normal, rendering the distribution of A'_i itself approximately normal, one cannot be sure.

We therefore carried out a number of numerical simulations based on the fictitious party system (2) and (3) (see Note 10). For each party system, we con-

No	**v**	**p**	β_2	β_3	**V**
(1)	$\left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)$	$\left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)$	−0.000	−0.693	$\begin{matrix}0.005 & \\ 0.003 & 0.008\end{matrix}$
(2)		$\left(\frac{11}{25}, \frac{28}{75}, \frac{14}{75}\right)$	−0.165	−0.856	$\begin{matrix}0.005 & \\ 0.002 & 0.008\end{matrix}$
(3)		$\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right)$	−0.406	−1.097	$\begin{matrix}0.005 & \\ 0.002 & 0.008\end{matrix}$
(4)		$\left(\frac{9}{25}, \frac{32}{75}, \frac{16}{75}\right)$	0.171	−0.525	$\begin{matrix}0.005 & \\ 0.003 & 0.007\end{matrix}$
(5)		$\left(\frac{3}{10}, \frac{7}{15}, \frac{7}{30}\right)$	0.443	−0.253	$\begin{matrix}0.005 & \\ 0.003 & 0.008\end{matrix}$
(6)		$\left(\frac{39}{100}, \frac{39}{100}, \frac{11}{50}\right)$	0.000	−0.573	$\begin{matrix}0.005 & \\ 0.003 & 0.007\end{matrix}$
(7)		$\left(\frac{3}{8}, \frac{3}{8}, \frac{1}{4}\right)$	0.000	−0.405	$\begin{matrix}0.005 & \\ 0.003 & 0.007\end{matrix}$
(8)		$\left(\frac{41}{100}, \frac{41}{100}, \frac{9}{50}\right)$	0.000	−0.823	$\begin{matrix}0.005 & \\ 0.002 & 0.008\end{matrix}$
(9)		$\left(\frac{17}{40}, \frac{17}{40}, \frac{3}{20}\right)$	−0.000	−1.041	$\begin{matrix}0.005 & \\ 0.002 & 0.009\end{matrix}$
(10)	$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$	$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$	−0.693	−0.693	$\begin{matrix}0.006 & \\ 0.002 & 0.006\end{matrix}$
(11)		$\left(\frac{11}{20}, \frac{9}{40}, \frac{9}{40}\right)$	−0.894	−0.894	$\begin{matrix}0.006 & \\ 0.002 & 0.006\end{matrix}$
(12)		$\left(\frac{5}{8}, \frac{3}{16}, \frac{3}{16}\right)$	−1.201	−1.207	$\begin{matrix}0.007 & \\ 0.002 & 0.007\end{matrix}$
(13)		$\left(\frac{9}{20}, \frac{11}{40}, \frac{11}{40}\right)$	−0.492	−0.492	$\begin{matrix}0.006 & \\ 0.002 & 0.006\end{matrix}$
(14)		$\left(\frac{3}{8}, \frac{5}{16}, \frac{5}{16}\right)$	−0.181	−0.184	$\begin{matrix}0.006 & \\ 0.003 & 0.006\end{matrix}$
(15)		$\left(\frac{29}{60}, \frac{29}{120}, \frac{11}{40}\right)$	−0.691	−0.563	$\begin{matrix}0.006 & \\ 0.002 & 0.006\end{matrix}$
(16)		$\left(\frac{11}{24}, \frac{11}{48}, \frac{5}{16}\right)$	−0.693	−0.381	$\begin{matrix}0.007 & \\ 0.002 & 0.005\end{matrix}$
(17)		$\left(\frac{31}{60}, \frac{31}{120}, \frac{9}{40}\right)$	−0.695	−0.832	$\begin{matrix}0.006 & \\ 0.002 & 0.006\end{matrix}$
(18)		$\left(\frac{13}{24}, \frac{13}{48}, \frac{3}{16}\right)$	−0.693	−1.064	$\begin{matrix}0.006 & \\ 0.002 & 0.007\end{matrix}$

Table 4: β_1, β_2, and **V** under 2×9 different types of bias

sidered nine different scenarios: no bias as well as under-/overrepresentation of support for the biggest/smallest party by factors of $\frac{1}{10}$ and $\frac{1}{4}$, respectively. The missing/superfluous support was taken/given to the other parties according to their true support in the population. Put differently, we simulated the more interesting case of *concentrated* bias, since diffuse bias comes close to no bias at all.

Table B lists the 2×9 different scenarios we were investigating, and the values for β_2 and β_3 which are implied by **p**, along with their variance-covariance matrix **V** for $n = 1000$. Under each scenario, we drew 50000 values from the joint distribution of β_2 and β_3 (given **V**), and calculated A'_1, A'_2 and A'_3. This resulted in 54 simulated sampling distributions for our measure. Given the very large number of simulated observations, we applied the skewness and kurtosis test to detect deviations from normality (D'Agostino, Belanger and D'Agostino, 1990).[18]

[18]Stata's `sktest` implements an adjustment of the original test suggest by Royston (1991).

Not withstanding the very large sample size, the test failed to reject the null hypothesis of normality (see Table B) in each instance. The lowest p-value for the omnibus test was 0.74. P-values for the component kurtosis test are somewhat smaller but do not go below 0.48. These results are further confirmed by a visual inspection of the quantile plots.

Moreover, the simulated sampling distributions are neatly centred on their expected theoretical value, i.e. there is no bias (see the leftmost column in each panel of Table B). As an additional safeguard, we ran another set of simulations using the same bias factors on two fictitious six party systems (5) and (6), resulting in a further 108 simulated sampling distributions of A'_i (not shown as a table). Again, none of the omnibus tests for normality has a p-value lower than 0.05, some of the component kurtosis tests do, hinting at tails that are slightly heavier than normal. In each case, however, the deviation is tiny, i.e. less than 0.05. We are therefore confident that A'_i is approximately normally distributed under a wide range of conditions, and that classical hypothesis tests and confidence intervals are valid.

This leaves the problem of finding a computationally efficient way to estimate the sampling variance of $A'_1, A'_2, \cdots A'_k$ for a given data set. Fortunately, using the delta method (Agresti, 2002, 73-74, ch. 14.1) it is easy to come up with good approximations of these quantities.

The delta method, whose foundations were laid in the 1940s by Cramér (Oehlert, 1992), approximates the expectation (or higher moments) of some function $g(\cdot)$ of a random variable X by relying on a (truncated) Taylor series expansion. More specifically, Agresti (2002, 578) shows that (under weak conditions) for some parameter θ that has an approximately normal sampling distribution with variance σ^2/n, the sampling distribution of $g(\theta)$ is also approximately normal with variance $[g'(\theta)]^2\sigma^2/n$, since $g(\cdot)$ is approximately linear in the neighbourhood of θ. The delta method can be generalised to the case of a multivariate normal random vector (Agresti, 2002, 579) such as the joint sampling distribution of some set of parameter estimates.

Stata's procedure `nlcom` is a particularly versatile and powerful implementation of the delta method. As a post-estimation command, `nlcom` accepts symbolic references to model parameters and computes sampling variances for their linear and non-linear combinations and transformations. Our add-on `surveybias` internally makes the required calls to `nlcom` in order to calculate approximate standard errors for the A'_is.

The approximation works very well, as can be gleaned from the rightmost column in each panel in Table B: in 54 experiments, the approximated standard error is never off by more than 0.2 per cent.

No	A'_1					A'_2					A'_3				
	$\mu - A'_1$	p_1	p_2	p_3	$\sigma/\hat{\sigma}$	$\mu - A'_2$	p_1	p_2	p_3	$\sigma/\hat{\sigma}$	$\mu - A'_3$	p_1	p_2	p_3	$\sigma/\hat{\sigma}$
(1)	−0.001	0.99	0.92	0.99	1.000	−0.001	0.63	0.63	0.79	1.000	−0.001	0.94	0.49	0.78	1.000
(2)	−0.001	0.98	0.90	0.99	1.000	−0.001	0.67	0.64	0.82	1.000	−0.001	0.94	0.49	0.79	1.000
(3)	−0.001	0.97	0.87	0.99	1.000	−0.001	0.72	0.67	0.86	1.000	−0.001	0.94	0.49	0.79	1.000
(4)	−0.001	1.00	0.94	1.00	1.000	−0.001	0.59	0.63	0.77	1.000	−0.001	0.94	0.49	0.78	1.000
(5)	−0.001	0.99	0.97	1.00	1.000	−0.001	0.53	0.65	0.74	1.000	−0.001	0.94	0.49	0.78	1.000
(6)	−0.001	0.98	0.90	0.99	1.000	−0.001	0.61	0.63	0.78	1.000	−0.001	0.94	0.49	0.78	1.000
(7)	−0.001	0.97	0.88	0.99	1.000	−0.001	0.59	0.63	0.77	1.000	−0.001	0.94	0.49	0.78	1.000
(8)	−0.001	1.00	0.93	1.00	1.000	−0.001	0.64	0.63	0.80	1.000	−0.001	0.94	0.49	0.79	1.000
(9)	−0.001	0.99	0.96	1.00	1.000	−0.001	0.67	0.65	0.82	1.000	−0.001	0.94	0.49	0.79	1.000
(10)	−0.001	0.89	0.71	0.92	1.000	−0.001	0.71	0.66	0.85	1.000	−0.001	0.94	0.49	0.78	1.000
(11)	−0.001	0.88	0.68	0.91	1.000	−0.001	0.75	0.69	0.88	1.000	−0.001	0.94	0.49	0.78	1.000
(12)	0.001	0.87	0.66	0.89	0.999	0.001	0.82	0.74	0.92	1.000	−0.006	0.94	0.49	0.79	0.998
(13)	−0.001	0.91	0.74	0.94	1.000	−0.001	0.66	0.64	0.81	1.000	−0.001	0.95	0.48	0.78	1.000
(14)	0.001	0.93	0.79	0.96	1.000	0.001	0.59	0.63	0.77	1.000	−0.004	0.95	0.48	0.78	0.999
(15)	−0.001	0.88	0.69	0.91	1.000	−0.001	0.69	0.65	0.83	1.000	−0.001	0.95	0.48	0.78	1.000
(16)	−0.001	0.86	0.66	0.89	1.000	−0.001	0.68	0.64	0.82	1.000	−0.001	0.95	0.48	0.78	1.000
(17)	−0.001	0.91	0.73	0.94	1.000	−0.001	0.72	0.67	0.86	1.000	−0.001	0.94	0.49	0.78	1.000
(18)	0.001	0.93	0.78	0.96	0.999	0.001	0.75	0.69	0.88	1.000	−0.006	0.94	0.49	0.79	0.998

p_1: test based on skewness; p_2: test based on kurtosis; p_3: χ^2 omnibus test

Variance and covariances for B and B_w are also approximated and posted by our add-on, although the respective distributions of these quantities are folded-normal and therefore skewed, as explained in section 2.4. Additionally, the full (approximate) matrix of parameter variances and covariances is stored for further post-estimation analyses.

C Statistical Power

The statistical power of the χ_2-test is a function of sample size, magnitude of bias, and distribution of bias amongst parties. To assess how badly a poll must be biased in practice to be flagged up by the test, we ran another series of simulations. For party systems (2), (3), (5), and (6), we simulated that support for the biggest/smallest party was deflated/inflated by a factor of 10, 15, and 25 per cent, respectively. Under each of these 48 scenarios, we sampled from the appropriate biased multinomial distribution. The sample size was set to 1000, which seems to be the lower limit for commercial polls. We then tested against the null hypothesis of no bias based on both Pearson's χ^2 and the likelihood ratio G^2, using the conventional criterion of $p \leq 0.05$. For each scenario, this procedure was carried out 10000 times.

Table 5 shows the results for the three-party systems. First, note that results are virtually identical for Pearson's χ^2 and the likelihood ratio G^2, and that the direction (upward/downward) does not matter (as it should). Second, even for moderately large values of B_w in the range of 0.12 to 0.15, the null hypothesis is rejected in the vast majority of cases, comfortably exceeding the conventional threshold of 0.8. If the total bias B_w is smaller than 0.1, however, the power of the test is considerably lower. If, for instance, support for smallest party in system (2) is systematically biased from 20 to 22 per cent in the polls, the test will miss this bias in roughly two out of three applications.

Our simulations for the six-party systems by and large confirm these findings (see Table 6). The rejection rates for moderate bias are somewhat lower yet still acceptable, while strong bias (B_w exceeding 0.2) is detected with certainty. Note, however, that even substantial bias in the measurement of support for small parties hardly affects B_w (under the somewhat unrealistic assumption that the resulting bias in the measurement of other parties is distributed proportionally), and is rarely picked up by the test (see e. g. the last line in Table 6). One should, however, remember that even this comparatively strong effect is equivalent of a measurement that is biased from 5 to 6.25 per cent. While the detection of bias

system	most affected party	bias	B_w	$\chi^2 \geq \chi^2_{crit}$	$G^2 \geq G^2_{crit}$
(2)	biggest	-0.25	0.33	1.00	1.00
(2)	biggest	-0.15	0.19	0.95	0.95
(2)	biggest	-0.10	0.13	0.64	0.64
(2)	biggest	0.10	0.13	0.63	0.63
(2)	biggest	0.15	0.19	0.94	0.94
(2)	biggest	0.25	0.32	1.00	1.00
(2)	smallest	-0.25	0.15	0.97	0.97
(2)	smallest	-0.15	0.09	0.57	0.56
(2)	smallest	-0.10	0.06	0.29	0.28
(2)	smallest	0.10	0.06	0.27	0.28
(2)	smallest	0.15	0.09	0.54	0.55
(2)	smallest	0.25	0.14	0.93	0.94
(3)	biggest	-0.25	0.41	1.00	1.00
(3)	biggest	-0.15	0.25	0.99	0.99
(3)	biggest	-0.10	0.16	0.82	0.82
(3)	biggest	0.10	0.17	0.82	0.81
(3)	biggest	0.15	0.25	0.99	0.99
(3)	biggest	0.25	0.44	1.00	1.00
(3)	smallest	-0.25	0.20	0.99	0.99
(3)	smallest	-0.15	0.12	0.72	0.70
(3)	smallest	-0.10	0.08	0.36	0.35
(3)	smallest	0.10	0.08	0.35	0.35
(3)	smallest	0.15	0.12	0.68	0.68
(3)	smallest	0.25	0.19	0.99	0.99

Table 5: Statistical Power: Three-Party Systems

system	most affected party	bias	B_w	$\chi^2 \geq \chi^2_{crit}$	$G^2 \geq G^2_{crit}$
(5)	biggest	-0.25	0.38	1.00	1.00
(5)	biggest	-0.15	0.23	0.97	0.98
(5)	biggest	-0.10	0.15	0.68	0.69
(5)	biggest	0.10	0.16	0.67	0.66
(5)	biggest	0.15	0.24	0.98	0.97
(5)	biggest	0.25	0.41	1.00	1.00
(5)	smallest	-0.25	0.07	0.54	0.50
(5)	smallest	-0.15	0.04	0.19	0.18
(5)	smallest	-0.10	0.03	0.11	0.10
(5)	smallest	0.10	0.03	0.10	0.11
(5)	smallest	0.15	0.04	0.18	0.19
(5)	smallest	0.25	0.07	0.46	0.49
(6)	biggest	-0.25	0.21	1.00	0.99
(6)	biggest	-0.15	0.12	0.69	0.68
(6)	biggest	-0.10	0.08	0.32	0.32
(6)	biggest	0.10	0.08	0.30	0.30
(6)	biggest	0.15	0.12	0.64	0.64
(6)	biggest	0.25	0.20	0.99	0.99
(6)	smallest	-0.25	0.03	0.25	0.22
(6)	smallest	-0.15	0.02	0.11	0.09
(6)	smallest	-0.10	0.01	0.07	0.07
(6)	smallest	0.10	0.01	0.08	0.08
(6)	smallest	0.15	0.02	0.10	0.11
(6)	smallest	0.25	0.03	0.23	0.26

Table 6: Statistical Power: Six-Party Systems

in highly fragmented party systems clearly requires larger samples, we are reasonably sure that our test has sufficient power to detect substantively relevant bias under a wide range of conditions.

References

Agresti, Alan. 2002. *Categorical Data Analysis*. 2 ed. Hoboken: John Wiley.

Aitchison, John. 1982. "The Statistical Analysis of Compositional Data (with Discussion)." *Journal of the Royal Statistical Society, Series B* 44(2):139–177.

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.

Bacon-Shone, John. 2011. A Short History of Compositional Data Analysis. In *Compositional Data Analysis. Theory and Applications*, ed. Vera Pawlowsky-Glahn and Antonella Buccianti. Chichester: Wiley pp. 1–11.

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):148–170.

Cheng, Simon and J. Scott Long. 2007. "Testing for IIA in the Multinomial Logit Model." *Sociological Methods & Research* 35(4):583–600.

D'Agostino, Ralph B., Albert Belanger and Ralph B. D'Agostino, Jr. 1990. "A Suggestion for Using Powerful and Informative Tests of Normality." *The American Statistician* 44(4):316–321.

Dow, Jay K. and James W. Endersby. 2004. "Multinomial Probit and Multinomial Logit. A Comparison of Choice Models for Voting Research." *Electoral Studies* 23:107–122.

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):275–298.

Evans, Jocelyn. 2012. "The Sound Foundations of a Socialist Victory." *Renewal* 20(2–3):123–128.

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):250–257.

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. **URL:** *http://www.unc.edu/ kropko/paper1.pdf*

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, London, New Delhi: 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):342–369.

McFadden, Daniel. 1973. Conditional Logit Analysis of Qualitative Choice Behaviour. In *Frontiers of Econometrics*, ed. Paul Zarembka. New York: Academic Press pp. 105–142.

McLean, Iain. 1995. "Independence of irrelevant alternatives before Arrow." *Mathematical Social Sciences* 30(2):107–126.

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.

Oehlert, Gary W. 1992. "A Note on the Delta Method." *The American Statistician* 46(1):27–29.

Pawlowsky-Glahn, Vera and Antonella Buccianti, eds. 2011. *Compositional Data Analysis. Theory and Applications*. Chichester: 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):987–991.

Royston, Patrick. 1991. "sg3.5: Comment on sg3.4 and an Improved D'Agostino Test." *Stata Technical Bulletin* 3:23–24.
URL: *http://stata-press.com/journals/stbcontents/stb3.pdf*

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 Prognosegüte von Wahlbörsen und Meinungsumfragen zur Bundestagswahl 2005." *Politische Vierteljahresschrift* 46(4):674–681.

Train, Kenneth. 2009. *Discrete Choice Methods with Simulation*. 2 ed. Cambridge: 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):231–260.