Australian Journal of Political Science,
Vol. 40, No. 4, December, pp. 499–517

Routledge
Taylor & Francis Group

Pooling the Polls Over an Election Campaign

SIMON JACKMAN

Stanford University

Poll results vary over the course of a campaign election and across polling organisations, making it difficult to track genuine changes in voter support. I present a statistical model that tracks changes in voter support over time by pooling the polls, and corrects for variation across polling organisations due to biases known as 'house effects'. The result is a less biased and more precise estimate of vote intentions than is possible from any one poll alone. I use five series of polls fielded over the 2004 Australian federal election campaign (ACNielsen, the ANU/ninemsn online poll, Galaxy, Newspoll, and Roy Morgan) to generate daily estimates of the Coalition's share of two-party preferred (2PP) and first preference vote intentions. Over the course of the campaign there is about a 4 percentage point swing to the Coalition in first preference vote share (and a smaller swing in 2PP terms), that begins prior to the formal announcement of the election, but is complete shortly after the leader debates. The ANU/ninemsn online poll and Morgan are found to have large and statistically significant biases, while, generally, the three phone polls have small and/or statistically insignificant biases, with ACNielsen and (in particular) Galaxy performing quite well in 2004.

Polls are the lifeblood of media coverage and punditry during an election campaign. Each major newspaper has a contractual arrangement with a polling organisation, providing them with poll numbers—grist for the journalistic mill, as it were—a critical component of the campaign coverage in the media. As an election draws closer, the tempo of polling increases, as does the number of polling organisations in the field, and the intensity with which the poll numbers are studied. In sum, media polls are perhaps the most readily available information on how the parties are faring over the campaign.[1]

Simon Jackman is an Associate Professor in the Department of Political at Stanford University and Director of the Political Science Computational Laboratory. An earlier version of this paper was prepared for the annual meeting of the Australasian Political Studies Association, University of Adelaide, 29 September–1 October 2004. I thank David Gow, Andrew Leigh, Kevin Quinn and Justin Wolfers for useful discussion, and Randall Thomas for some useful references.

[1] Other sources of information include the prices offered in election betting markets (eg Wolfers and Leigh 2002) but my focus here is strictly on the polls.

ISSN 1036-1146 print; ISSN 1363-030X online/05/040499-19
DOI: 10.1080/10361140500302472

Although media-commissioned polls are central to the media's coverage of the election campaign, they are actually of limited use for political scientists. As I show below, media-commissioned polls employ sample sizes that are too small to reliably detect the relatively small day-to-day or week-to-week movements in voter sentiment we would expect to occur over an election campaign. Thus, as they currently stand, media-commissioned polls are essentially of 'news value' rather than social-scientific value. As Murray Goot explains:

> The problem in the reporting of the polls does not arise, fundamentally, from the fact that journalists are ill-trained to deal with such data—though that to some extent is true. Fundamentally, the press plays up differences which are otherwise insignificant because it *has* to. Its only alternative is to say that what a poll found today is not significantly different from what it found yesterday; and under most (though not all) circumstances, that sort of news is no news at all. (Goot 2000, 46)

Likewise, in commenting on the polls in the 2004 election, Peter Brent (2004) observes that

> [q]uantitative opinion polls aren't that precise. But the process that pays for them pretends they are. They [polls] cost a bundle and so are given pride of place. Once they're there, everyone involved goes along with the charade.

Beyond endorsing these observations, my goal is not to dwell on or further criticise the media's reporting of poll results. Rather, my goal here is more constructive, to present tools that make better use of published polls, so as to more reliably track changes in voter sentiment over the course of an election campaign. In short, while the precision of any one poll is quite limited, I show that we can systematically combine the information in the published polls, leveraging them against one another, so as to obtain a clearer picture of what might be going on in the electorate over the campaign.

In the next section I discuss the limits of polls, highlighting (1) imprecision due to sampling error; and (2) the possibility of bias, due to the procedures and methods employed by the particular polling organisation. Pooling the polls helps deal with the first problem (more data are better than less), but is a valid strategy only if polls are all measuring the same thing (ie each poll produces an unbiased estimate of the current state of voter support). In addition, (3), voter support is likely to be changing over the course of a campaign, while polls provide snapshots, and imprecise and possibly biased snapshots. I then present a statistical model that addresses these three issues simultaneously: ie (1) the model *pools* the polls, overcoming the limits to precision inherent in any one poll; (2) the model smoothes over time, consistent with the notion that although support for either party fluctuates or trends over the course of a campaign, each campaign day need not be considered *de novo*—yesterday's level of Coalition support will generally be an excellent predictor of today's level of support; (3) the model estimates and corrects for the possibility that any single poll is subject to bias or 'house effects' (bias induced by methodological procedures specific to each polling organisation). I apply this statistical model to polling data generated in the lead-up to the 2004 Australian federal election, drawing on data from five polling organisations, spanning traditional telephone interviewing through to face-to-face and self-selected samples completing surveys via the Internet.

The chief benefit of pooling poll results (after correcting for house effects) is that we are much better positioned to ascertain movements in levels of voter support in response to campaign events. As it turns out, there is simply not a tremendous amount of volatility over the course of the 2004 campaign, at least as far as we can detect it with the available polling data and the model I use here: to foreshadow one of my findings, in terms of aggregate, two-party preferred (2PP) voting intentions, there was a steady shift towards the Coalition over the opening weeks of the campaign, and little movement thereafter. Larger movements are apparent in the Coalition's share of first preferences over the course of the campaign. But a striking feature of the 2004 polling data is that a good portion of the differences across polls is not due to movements in voter sentiment but to differences in the polls' methodologies. The model and analysis I present here lets us distinguish between these different sources of variation in polling data: movement due to fluctuations over the campaign, sampling error, and non-sampling error (bias) specific to each polling organisation ('house effects').

The Limits of Polls: Margins of Error and Sample Size

A poll's reported 'margin of error' is almost always a 95% confidence interval around the poll's estimate of the 2PP vote split, provided by the approximation (valid for large samples)

$$\hat{\alpha} \pm 1.96\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{n}}, \tag{1}$$

where $\hat{\alpha} \in [0,1]$ is the poll estimate of α, some proportion of interest (eg the proportion of 2PP vote intentions for a particular party) and n is the sample size. The 1.96 in equation (1) comes from the fact that (1) via the central limit theorem, uncertainty about a statistic such as $\hat{\alpha}$ computed with a large sample follows a normal distribution, and (2) if a random variable follows a normal distribution, then, with probability 0.95, it lies within 1.96 standard deviations of the mean of that normal distribution.

Equation (1) highlights the fact that statistical precision increases in the square root of the sample size, so that increases in sample size produce diminishing marginal gains in statistical precision. But the marginal cost of an additional survey respondent remains more or less fixed, or may even be increasing in sample size. Since time and money are not infinite, there are limits to the precision we can reasonably expect from any given poll. Most media polls are conducted with sample sizes in the neighbourhood of 1000–1500, with larger samples of above 2500 being fielded in the days immediately prior to the election. These sample sizes are simply too small to reliably detect small fluctuations in support for the parties over the course of an election campaign. Routine statistical calculations[2] reveal that sample sizes of the order of 65,000 per sample are required to give a researcher a 95% chance of detecting a true 1 percentage point change with a conventional 95% level of confidence; if the researcher is more risk acceptant and is willing to accept a lower level of statistical significance

[2] See, for example, Fleiss, Levin and Paik (2003, ch. 4). For the specific context of sample sizes required to assess campaign effects, see Gow (2001).

the sample size requirements become slightly less onerous, but still well beyond the resources deployed in media-commissioned polls. For instance, to give a researcher an even-money chance of detecting a 1 percentage point change using a 90% confidence level, the researcher requires roughly 13,500 respondents per sample.

Larger changes in voter support are easier to detect, in the sense of requiring fewer respondents. Figure 1 summarises the sample size requirements over a range of assumed differences between polls. For instance, around 4000 respondents per poll are sufficient to detect changes of 4 percentage points (again, in the sense that a researcher has a 95% chance of rejecting the null hypothesis of no change at the traditional benchmark of 95% confidence levels); for changes this large, a sample size of 1200 per week is sufficient to give the researcher a 50–50 chance of rejecting the null hypothesis of no change at the 95% confidence level.

To summarise, the sample sizes used by the published media polls (eg the 1400 respondents typically seen in the ACNielsen polls for the Fairfax papers) have a reasonable chance of detecting moderate to large swings in voter support. But the probability of detecting a small swing, say, of the order of a percentage point, is less than 1 in 10, given the sample sizes used by most commercial polls. To the extent that campaigns generate these small to moderate changes in voter support, we clearly need more data than those provided by a single media-commissioned poll if we are to detect it.

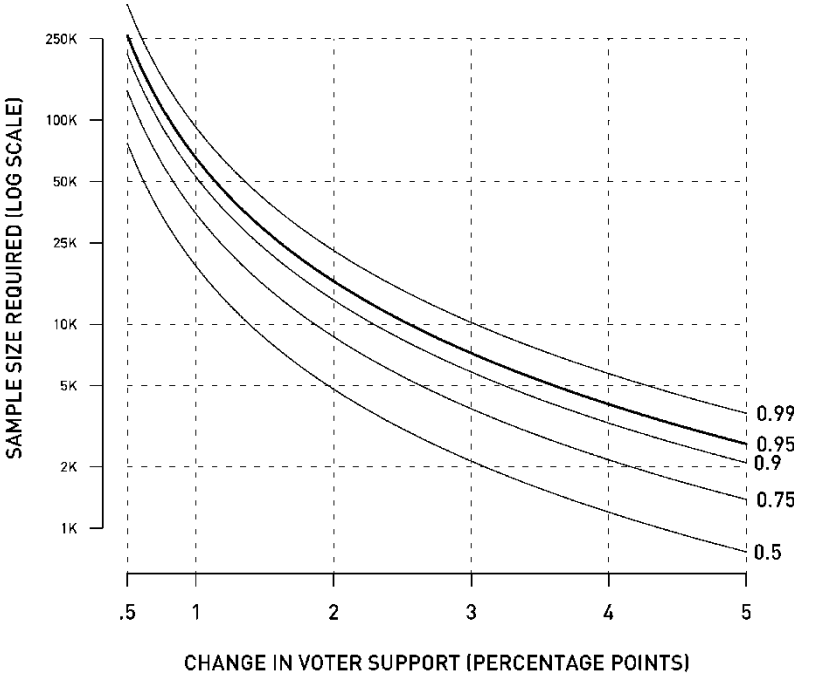

Figure 1. Sample size requirements. *Notes*: Each curve shows the sample size (vertical axis, log scale) required to detect the indicated change in support (horizontal axis, assuming a baseline level of 50%), with probability given by the label next to each line. In each instance it is assumed that the researcher's decision problem is whether to reject the null hypothesis of no change in favour of a two-sided, alternative hypothesis, using a 95% confidence level or better (ie a p-value of 0.05).

Pooling the Polls

One way to get more data is to *pool* the polls. If two or more polls are in the field at more or less the same time, then we have the potential to combine the information in each poll, to arrive at an estimate that is more precise than any single poll. Suppose that two polling organisations, A and B, generate unbiased estimates of α, the Coalition's 2PP vote share. Organisation A conducts a poll with sample size n_A and produces the estimate $\hat{\alpha}_A$, and so, recalling equation (1), our beliefs about α given A's poll can be represented with a normal distribution with mean $\hat{\alpha}_A$ and standard deviation $s_A = \sqrt{\hat{\alpha}_A(1-\hat{\alpha}_A)/n_A}$. Likewise, we obtain $\hat{\alpha}_B$ and s_B from organisation B. Since A and B are trying to estimate the same quantity, we can combine the information in their polls; in particular, it is a fact about normal distributions that the pooled estimate is a *precision-weighted average* of the two separate estimates: ie

$$\hat{\alpha}_{AB} = \frac{p_A\hat{\alpha}_A + p_B\hat{\alpha}_B}{p_A + p_B}, \tag{2}$$

where p_A and p_B are the *precisions* of the two polls, defined as $1/s_A^2$ and $1/s_B^2$, respectively. The standard deviation we obtain using the pooled estimate is $\sqrt{1/(p_A+p_B)}$, and so the pooled estimate is more precise (has smaller standard deviation) than any one of the polls individually.

Pooling: An Example

Figure 2 presents a graphical illustration of the potential gains from pooling. Suppose polling company A estimates the Coalition's 2PP vote share at 53%, based on a sample size of 1400 respondents, and so via equation (1) the poll has a 'margin of error' of 2.6 percentage points. Suppose further that polling company B is in the field at roughly the same time and estimates the Coalition's 2PP vote share at 50% based on a sample size of 2500 respondents. Because of the larger sample size, company B's estimate has a smaller margin of error than the estimate of company A, of 2.0 percentage points. When pooling the two estimates, the greater precision of company B's estimate means we give more weight to it. Using equation (2), the pooled estimate is the precision-weighted average, 51.1%, slightly closer to B's estimate of 50% than the simple unweighted average of 51.5%. The pooled estimate has a margin of error of 1.5 percentage points, smaller than the margins of error associated with either A or B's results taken separately; after pooling the results, we conclude that with 95% probability, the Coalition's 2PP vote share lies between 49.5% and 52.6%.

'House Effects': Bias in the Polls

Pooling polls will always enhance precision, but the validity of the pooled estimate rests on a critical assumption: that the polls are unbiased. This is often not the case. Polls are subject to bias, and, in particular, biases specific to particular polling organisations, known as 'house effects' (the term 'house' here refers to a polling company, not a sampled household). Variations in mode of interview (telephone, face-to-face, or Internet), sampling and weighting procedures, the day of the week

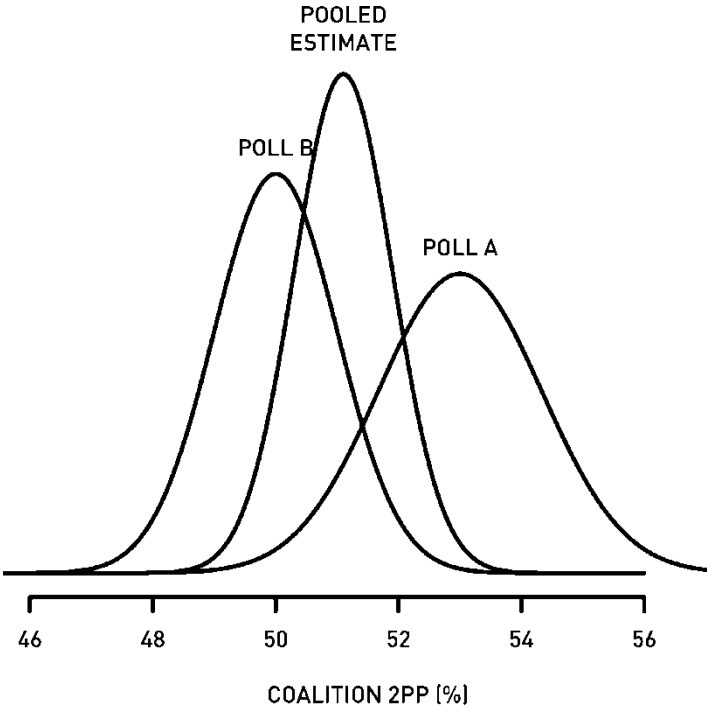

Figure 2. Pooling the polls: an example. *Notes*: The uncertainty attaching to each poll's estimate can be summarised with a normal distribution; see equation (1). Pooling the information in polls *A* and *B* produces a normal distribution that has (a) its mean equal to a precision-weighted average of the two polls, via equation (2), and (b) a smaller standard deviation than the normal distributions summarising polls *A* and *B*, reflecting the fact that pooling is equivalent to combining information.

or the time of day a company interviews, the age, ethnic and gender composition of a given company's interviewers, question wording and question ordering are all potential sources of bias in survey research. A critical facet of polling in the Australian context is how or whether to solicit second preferences from respondents who report a first preference for a minor party; as we shall see below, differences on this score can generate quite different estimates of 2PP vote shares (see also Brent, 2004). All survey companies make choices about these aspects of surveying, balancing considerations of cost and timeliness against beliefs about the magnitude of the biases likely to be introduced from the sources listed above; Groves (1989) provides a book-length survey and McDermott and Frankovic (2003) examine the magnitudes of specific sources of house effects in pre-election polling for the 2000 US presidential election. Again, for commercial reasons, survey companies tend to settle on a set of procedures (mode of interview, sampling, instrumentation, weighting); thus, any biases in one poll from a given organisation are often present in their other polls, and a more or less fixed aspect of their estimates. But the consequence of house effects is that sampling variation is not the only source of error. Put differently, the biases in poll *A* and poll *B* may be such that they can't validly be considered to be estimating the same population quantity, and pooling them would be invalid.

On the other hand, there are commercial pressures to eradicate bias, and survey research firms will review and occasionally alter their procedures to this end.

Election polling is somewhat unusual in this regard, since the election outcome itself provides a highly visible benchmark against which to assess a poll's estimates and, by extension, the validity of the polling organisation's methodology; see, for example, the essays by Goot (2000, 2002, 2005) reviewing the performance of the polls over recent Australian election campaigns. It is rare (but not unheard of) for polling companies to change their procedures in the middle of a high-profile event like a national election campaign;[3] acting on the advice of an anonymous referee, I contacted each of the polling organisations whose data I draw on here to verify that their procedures remained constant over the course of the campaign (and I received assurances that this was the case).

In summary, naively pooling the polls addresses one issue (the relative lack of precision in any one poll), while assuming house effects are non-existent. In general, this is a dangerous strategy: pooling two biased polls does not necessarily remove bias, since the pooled estimate will inherit some of the biases of each poll (in proportion to the precision of each poll). If the biases offset (say, poll *A* has a pro-Coalition bias and poll *B* has an ALP bias of roughly the same magnitude) then pooling will tend to alleviate the bias of any one poll. But if the biases run in the same direction, then the gain in precision from pooling results in higher levels of confidence about a biased estimate.

This indicates that we require estimates of the bias in each poll as we attempt to pool the polls. Of course, presuming that we have knowledge of each poll's biases is equivalent to knowing the population quantity being estimated. This is rarely ever the case, at least before an election: indeed, if we knew the population quantity of interest (eg the Coalition's share of 2PP vote intentions), we wouldn't be conducting polls! After the election, the actual election outcome provides a critical reference point, from which we can calibrate the various polls and, in turn, estimate house effects. I exploit this strategy in the analysis below.[4]

The Polls in 2004

I consider the following major national polls:

(1) ACNielsen, which provided polls for the Fairfax papers (eg the *Sydney Morning Herald* and *The Age*);
(2) Newspoll, which provided polls for News Ltd, primarily its national daily *The Australian*;
(3) Galaxy, which provided polls for other News Ltd metropolitan daily newspapers, such as Sydney's *Daily Telegraph* and the Brisbane *Courier Mail*;
(4) Roy Morgan, apparently without a media client in 2004, but formerly providing polls to *The Bulletin*;

[3] In the 2000 US presidential election campaign, some large fluctuations in poll estimates were due to changes in weighting procedures. Another complication and likely source of bias in the United States is the definition of a likely voter; these definitions vary across organisations and sometimes change within organisations over the course of a campaign. See Erikson, Panagopoulos and Wlezien (2004).

[4] Of course, this is not possible before the election: possible strategies for dealing with house effects over the course of the campaign include (a) using past assessments of the performance of the polls (although this isn't feasible for new polling companies, with no history of election polling); (b) imposing an assumption of some sort, such that the polls are collectively unbiased, such that, on average, the polls get it right (as I did in a 'pre-election' version of this paper; as I show below, this assumption is not supported by the data).

(5) the Australian National University, which provided daily polls from 12 September up until 8 October (election eve) for Consolidated Press Holdings outlets such as *ninemsn.com*, Channel Nine, and *The Bulletin*.

Three of the five organisations use telephone polling, the exceptions being Morgan (almost exclusively face to face) and the ANU/*ninemsn.com* collaboration (polling via the Internet).

Between mid-June 2004 and the election on 9 October, these five organisations published the results of 66 polls, comprising over 72,000 interviews via phone (the dominant interview mode), face to face, and Internet. The tempo of polling increased markedly after the formal announcement of the election. The data are summarised graphically in Figure 3. In the figure, each poll's estimate of the Coalition's share of the 2PP vote (vertical axis) is plotted over time with the mid-point of a poll's field period taken as the effective date of the poll. The vertical lines in the figure cover 95% confidence intervals around each poll's estimate of the Coalition's share of 2PP vote intentions, using the formula in equation (1). In the analysis, below, I also examine the polls' estimates of the Coalition's first preference vote share.

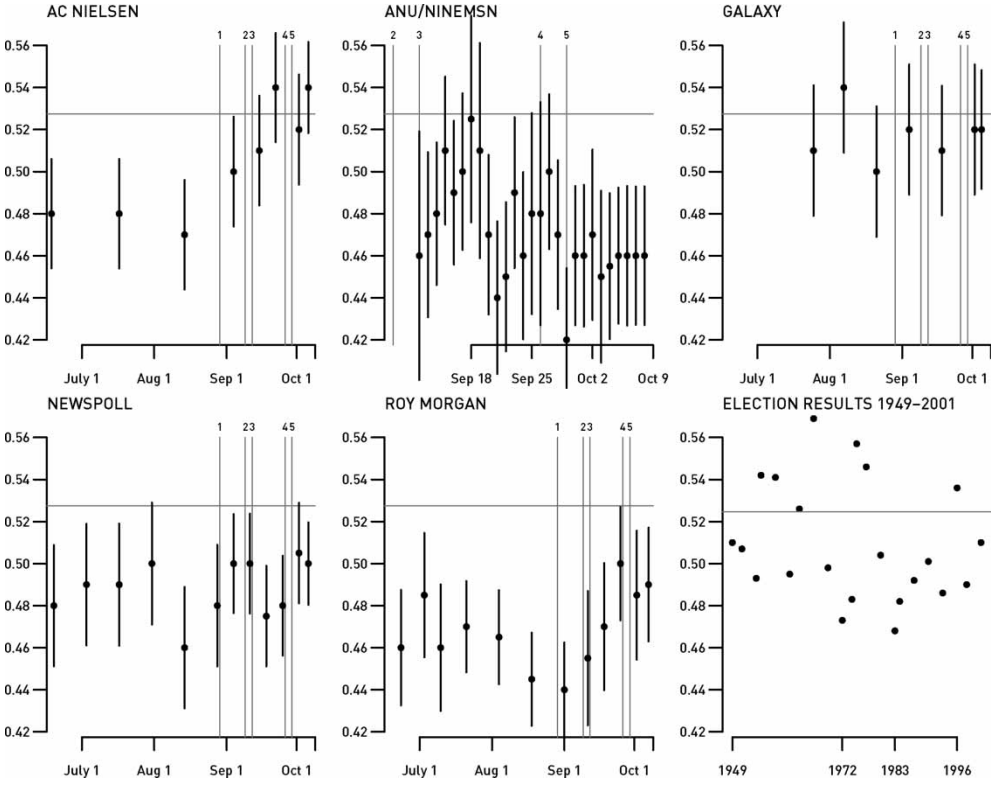

Figure 3. Polls in the 2004 campaign and historical election outcomes. *Notes*: Each poll is represented by a dot, with the vertical lines extending to cover a 95% confidence interval. The vertical axis is identically scaled in each panel; note that the ANU/ninemsn series of daily polls starts during the campaign. The vertical lines labelled 1 to 5 correspond to the following campaign events: (1) election announced, 29/8; (2) Jakarta Embassy bombing, 9/9; (3) Leader debate, 9/12; (4) Liberal Party campaign launch, 9/26; (5) ALP launch, 9/29.

The Evidence for House Effects in 2004

Figure 3 makes several features of the 2004 polls quite obvious. First, most of the polls do not provide unambiguous evidence that the Coalition is leading the ALP or vice versa. Telling exceptions are: (1) most of the Morgan polls, especially those from July, August and early September, with relatively large sample sizes (and hence smaller confidence intervals) that strongly suggest a sizeable Labor lead; (2) a mid-August Nielsen poll that found Labor in the lead, and the 23–24 September and the election eve Nielsen polls that estimated the Coalition's 2PP vote share at 54%; (3) two Newspolls that again point to a Labor lead; (4) 12 of the ANU/ninemsn daily Internet polls strongly indicate a Labor lead; and (5) the early August Galaxy polls showing the Coalition unambiguously leading the ALP. Tellingly, the polls that do give an unambiguous lead to one party or the other point in contradictory directions. In short, the polls are all over the map: some polls point to a Coalition win with high probability; some point to a Labor win with high probability; and most polls fall somewhere in between.

To give some sense of the variability across the polls, the lower right panel of Figure 3 shows the distribution of actual results in the 22 House of Representatives elections from 1949 to 2001. The Coalition's 2PP vote share has never been below the 46.8% result recorded in 1983 and has never exceeded the 56.9% result recorded in 1966, a range of 10.1 percentage points. Half of the Coalition's 2PP results lie in a range spanning 49.1–53.4%, or 4.2 percentage points. In contrast, the polls in the second half of 2004 vary between a low of 42% (recorded in the ANU/ninemsn online poll of 29 September) and a high of 54% (recorded in a Galaxy poll from early August, and two Nielsen polls, including the Nielsen poll immediately before the election). That is, the range of poll results is larger than the range observed in 22 actual elections covering some 55 years of Australian political history.

Pursuing this point, note that the standard deviation of the 22 elections (1949–2001) is 2.8 percentage points. This degree of variation roughly corresponds to the uncertainty that accompanies a poll with a sample size of about 315 respondents.[5] How is it that polls using sample sizes typically in the neighbourhood of 1000 respondents display essentially just as much variability as the (nominally less precise) historical data?

The data presented in Figure 3 provide a strong hint that 'house effects' are at work in the 2004 polling data. The polls' point estimates of Coalition vote share are widely dispersed, to be sure; but the variation across polling organisations in the point estimates is substantial. That is, large fluctuations within the data from any single polling organisation are rare; eg the ANU/ninemsn estimates display the greatest variation, in no small measure because they are based on relatively small daily samples (an average daily sample size of 685). The Morgan polls seem to be consistently on the low side of the estimates of the Coalition's 2PP, while Galaxy seems to be consistently on the higher side of the estimates. In fact, simple analysis of variance (ANOVA) reveals that 41% of the variation in the polls' point estimates is house-to-house variation; an F-test leads to an overwhelming rejection of the null hypothesis that there are no differences among the polling organisations ($F_{4,61} = 10.61,\ p = 0.001$). Augmenting the analysis with various trend terms

[5] That is, if a poll of 315 respondents generated an estimate of the Coalition's vote share of 50%, then via equation (1), a 95% confidence interval would be plus or minus 1.96 times $\sqrt{0.25/315}$ or about 1.96 times 0.028.

(to address the hypothesis that the house effects are confounded with possibly real fluctuations in vote support) generates identical results.[6] The extent of this house-to-house variation is even more impressive when we recall that sampling error alone is large and an unavoidable source of variation across the polls; ie even without house effects and/or change in voter sentiment over the campaign, sampling variation alone would generate considerable dispersion in the polls' point estimates. The model presented below provides a more rigorous assessment of house effects, by explicitly dealing with both sampling error and the possibility that voter sentiment is moving over the campaign period.

A Statistical Model for Pooling the Polls

Here I present a statistical model that simultaneously tackles the three problems discussed above: (1) pooling polls so as to increase precision; (2) estimating and adjusting for the bias of any one poll; (3) tracking the trends and fluctuations in voter sentiment over the course of the election campaign. Some notation will help to clarify matters. Let α_t be the Coalition 2PP intended vote share at time t, with t indexing days, where $t = 1$ on 18 June 2004 (corresponding to the field date of the first poll in my data set); below, I also consider the polls' estimates of the Coalition's share of first preference votes. Let $i = 1, \ldots, n$ index the polls available for analysis. Each poll result is assumed to be generated as follows:

$$y_i \sim N(\mu_i, \sigma_i^2), \tag{3}$$

where y_i is the result of poll i. Each of the n polls is generated by organisation j_i on field date t_i. σ_i is the standard error of the poll (a function of y_i and the poll's sample size; again, see equation (1)) and

$$\mu_i = \alpha_{t_i} + \delta_{j_i}, \tag{4}$$

where δ_j is the bias of polling organisation j, an unknown parameter to be estimated.

To model change in vote intentions, I use the following simple random-walk model:

$$\alpha_t \sim N(\alpha_{t-1}, \omega^2),\ t = 2, \ldots, T \tag{5}$$

with the distribution

$$\alpha_1 \sim \text{Uniform}(0.4, 0.6) \tag{6}$$

initialising the random walk (ie before we see any polling, I assume that Coalition support is anywhere between 40% and 60%, bracketing the historical range of election results reported above). In adopting this model I assume that vote shares are *locally constant*, ie on average, today's level of Coalition support is the same

[6] For instance, adding a cubic polynomial in time to the analysis picks up an extra 0.03 in r^2; the r^2 from the regression with the cubic trend term alone is just 0.01.

as yesterday's, save for random shocks that come from a normal distribution with mean zero and standard deviation ω.

The model is fit subject to the constraint implied by the actual election outcome; that is, on 9 October 2004, $\alpha_t = 0.5274$ (ie the Coalition received 52.74% of the 2PP House of Representatives vote, according to the Australian Electoral Commission). This constrains the trajectory of the estimated α_t to culminate in the actual election outcome, and, in turn, lets us estimate the δ_j parameters tapping house-specific biases. This is an important constraint; without being able to anchor the estimated levels of Coalition support to the actual election outcome, the model unravels, it being impossible to simultaneously estimate underlying levels of support for the Coalition and house effects. Accordingly, the model cannot be used for 'real-time' tracking over the course of the campaign unless we impose a priori restrictions on the house effects; in fact, one goal of the current analysis is to provide estimates of house effects that then might be used to calibrate poll results in the *next* election campaign.

This model is essentially a Kalman filter, used in signal processing in engineering applications, for tracking a moving target with noisy (and possibly biased) observations. Kalman filters are also used widely in time-series econometrics (eg Harvey, 1989) and in political science, where they have been used to track presidential approval (eg Beck 1990; Baum and Kernell 2001) and public opinion (eg Stimson 1991; Green, Gerber and De Boef 1999). Equations (3) and (4) define the measurement or observational part of the model, relating the observed poll estimate to the latent target α_t, while equation (5) specifies the way in which the hidden target moves over time. The unknown model parameters are: (a) 114 α_t parameters, the daily levels of intended 2PP vote share for the Coalition over the 114 'campaign days' in my analysis; (b) the five bias parameters δ_j specific to each polling company; and (c) ω, the standard deviation of the normal distribution characterising day-to-day volatility in the α_t. The data used to estimate these parameters are the survey results y_i and their standard errors s_i. Note that not all polling organisations report data on every day; in fact, I have data from just 66 polls. At first glance, the paucity of data relative to the number of parameters would seem to prohibit meaningful statistical analysis, but this is misleading: in the absence of polling data, equation (5) provides a predictive model for α_t, but constrained by the influence of previous polling data on estimates of $\alpha_{t-1}, \alpha_{t-2}, \ldots$, and of future polling data on estimates of $\alpha_{t+1}, \alpha_{t+2}, \ldots$. Estimation of the model is via Bayesian simulation methods discussed in the Appendix. All computer programs and data are available upon request.

Results

Daily estimates of the Coalition's share of 2PP vote intentions are shown in Figure 4. The solid line connects the 114 daily estimates of the α_t parameters, and the dotted lines indicate a 95% confidence interval around each daily estimate. As in Figure 3, campaign events are indicated with vertical lines, and actual poll results are overlaid with plotted points.

There is some evidence of trending over the campaign in the data, with a small rise in the Coalition's share of the 2PP vote after the election was formally announced on 29 August 2004. For this date, the estimate of the Coalition's share of the intended 2PP vote is 50.6%, with a 95% confidence interval ranging from 48.2% to 52.6%. The final election result of 52.7% lies just at the upper limit of the 95% confidence interval of Coalition support at the formal start of the campaign; put differently, the

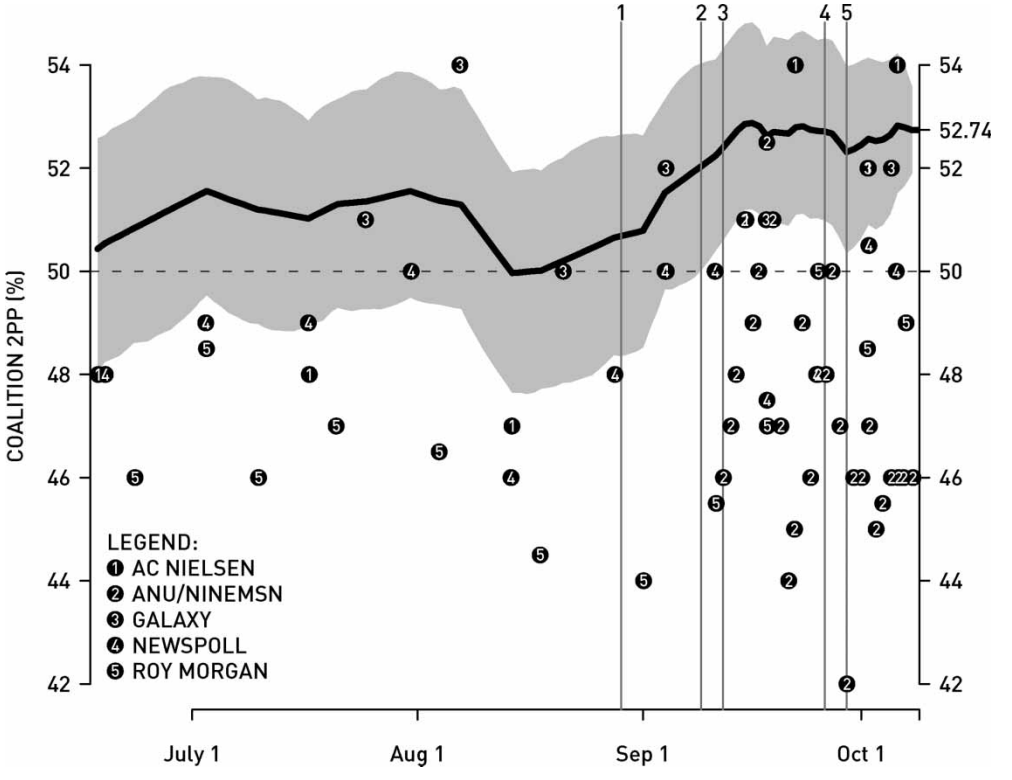

Figure 4. Estimated Coalition share of two-party preferred vote intentions, and pointwise 95% confidence intervals. *Notes*: The shaded area covers the 95% confidence intervals around the estimated levels of Coalition support, given the model and the polls. Individual polls are represented with a plotted point at their respective point estimates. See Figure 3 for campaign events.

probability that the Coalition gained support over the campaign is 0.977 (ie we would reject the null hypothesis of 'no change over the campaign' with greater than the conventional standard of 95% confidence).

My estimates also indicate that Coalition 2PP support peaked on 16 September, some three weeks before the election (but just after the leader debates) at 52.8% of the intended 2PP. Of course, there is considerable uncertainty accompanying this point estimate; a 95% confidence interval ranges from 51.0% to 54.5%, indicating that this 'peak' estimate is indistinguishable from the actual election outcome (52.7%). In sum, the estimates reported here indicate that the Coalition's gains over the campaign amounted to about 2 percentage points, almost the size of the overall swing to the government in the 2004 election, and that they came relatively early in the campaign, as events such as the Jakarta embassy bombing and the leaders' debate took place. The formal policy launches of both major parties seem to have made no impact on aggregate, 2PP vote shares.

The benefits of pooling data are also apparent from Figure 4. The width of the shaded area—a 95% confidence interval around the estimated α_t, the Coalition share of 2PP vote intentions—is substantially smaller than the confidence intervals attaching to any one poll. Moreover, as Figure 5 highlights, the width of the 95% interval is falling over time as more polls are published in the closing weeks of

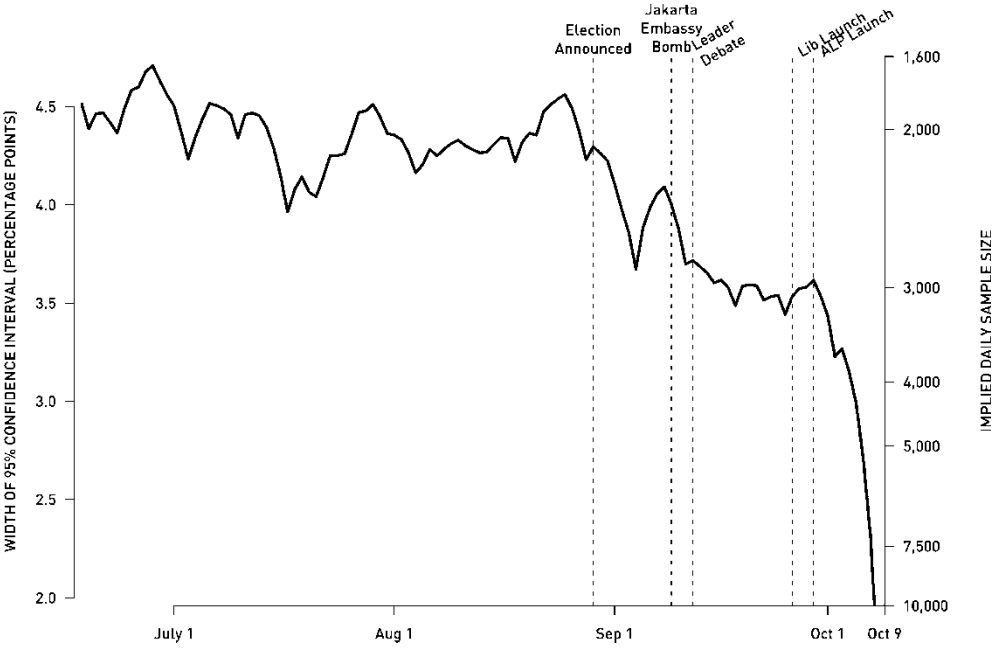

Figure 5. Width of 95% confidence interval around daily estimates of intended Coalition 2PP vote share. *Notes*: The axis on the right-hand side of the graph shows the implied sample size of a hypothetical poll if it were to generate a confidence interval as precise as that obtained by the pooling model.

the campaign. In particular, the constraint that estimated series of α_t culminate with the actual election result on 9 October sees the confidence interval shrink towards zero (and the implied sample size tend to infinity) in the last days of the campaign. A more realistic appraisal of the benefits of polling is the pattern, say, over the second last week of the campaign, where polling is quite frequent and the benefits of pooling and smoothing are quite substantial: in this time period the 95% confidence intervals are never more than 3.5 percentage points in width, corresponding to daily polls with sample sizes of the order of 3000 respondents.

A similar picture emerges when we turn to consider the polls' estimate of first preferences. Figure 6 presents the estimated daily track of first preference vote shares, generated by applying the model presented above to the poll data on Coalition first preferences. The Coalition's actual share of the first preference vote, 46.7%, is statistically distinguishable from the Coalition share of first preference vote intentions as estimated at the start of the series (43.5%, on 18 June, $p < 0.01$), at the estimated low point in Coalition support on 14 August (42.3%, $p < 0.01$), and on the day the election was formally announced (44.0% on 29 August, $p < 0.01$; all one-sided p-values). On the other hand, the Coalition's maximum level of estimated support (47.2%) is not statistically distinguishable from the actual election result of 46.7% ($p = 0.29$, one-tailed). Thus, the overall pattern is extremely similar to that which we observe for the 2PP estimates (see Figure 4): ie a shift towards the Coalition that appears to pre-date the formal announcement of the election, levelling off after the leader debate, and with no

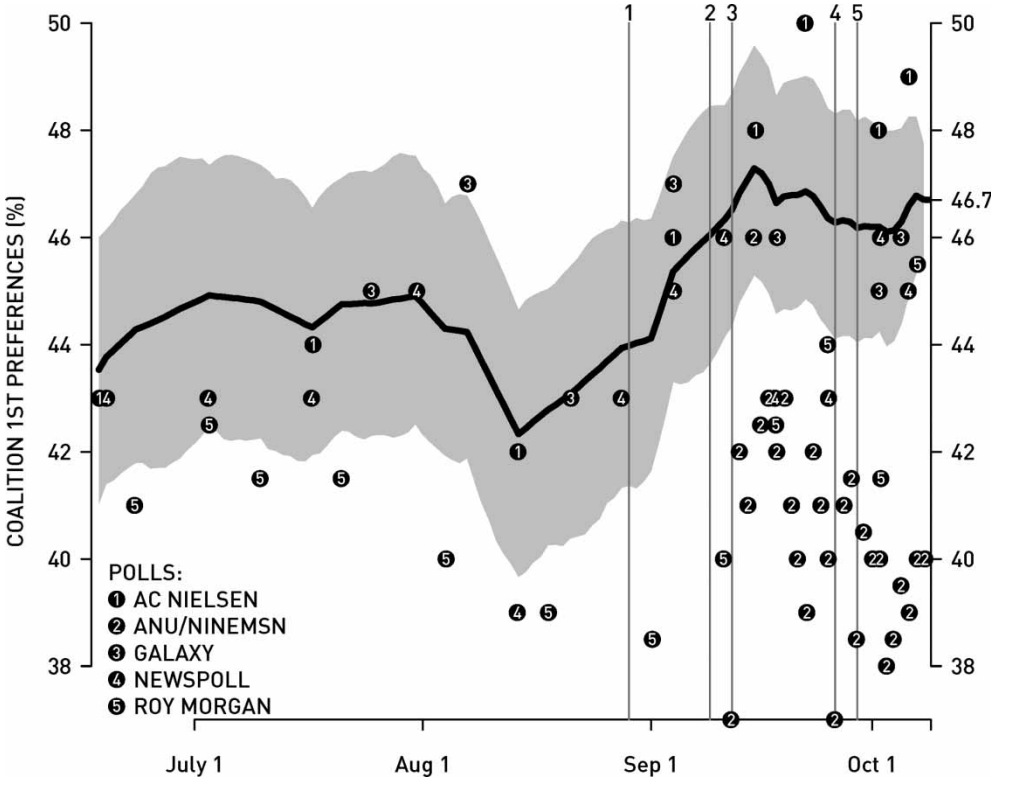

Figure 6. Estimated Coalition share of first preference vote intentions and pointwise 95% confidence intervals. *Notes*: The shaded area covers the 95% confidence intervals around the estimated levels of Coalition support, given the model and the polls. Individual polls are represented with a plotted point at their respective point estimates. See Figure 3 for campaign events.

statistically discernible movement thereafter. This movement towards the Coalition in terms of first preferences is over 4 percentage points, a large and politically consequential shift by any reasonable standard,[7] and is considerably larger than the movement in 2PP vote intentions.

Estimates of House Effects

Table 1 summarises the estimates of house effects, presenting point estimates and the lower and upper limits of a 95% confidence interval around the point estimates (ie point estimates with 95% confidence intervals overlapping zero are indistinguishable from zero at the traditional 95% confidence level). Results are provided for both the 2PP data and the polls' estimates of first preference vote shares.

For the 2PP estimates, three out of five organisations have estimated bias parameters that are distinguishable from zero at conventional levels of statistical

[7] Consider that the 4.4 percentage point movement from the Coalition's lowest level of estimated first preference support (42.3% on 14 August) to the actual result of 46.7% corresponds to the difference between a clear victory (ie the 2004 result) and a defeat (1972, 41.5%) or the slimmest of wins (1961, 42.0%).

Table 1. Estimates of house effects parameters, and limits of 95% confidence intervals. Positive house effects imply a systematic overestimate of Coalition support. All terms are expressed as percentage points

	Two-party preferred			First preferences		
	Estimate	2.5%	97.5%	Estimate	2.5%	97.5%
AC Nielsen	−1.1	−2.8	0.6	1.1	−0.8	3.0
ANU/ninemsn	−5.7	−7.3	−4.2	−6.0	−7.7	−4.2
Galaxy	0.1	−2.0	1.8	0.4	−1.6	2.5
Newspoll	−2.9	−4.6	−1.2	−1.4	−3.3	0.4
Roy Morgan	−4.9	−6.6	−3.2	−3.7	−5.5	−1.8
Average	−2.8	−4.3	−1.2	−2.0	−3.7	−0.3
Average of Nielsen/ Galaxy/Newspoll	−1.3	−2.9	0.2	−0.1	−1.9	1.8

significance, and all tending to underestimate Coalition support. The ANU/ninemsn online poll has the largest bias, underestimating Coalition support by an average of 5.7 percentage points; to better appreciate the magnitude of this bias, recall that between 1949 and 2004 actual Coalition 2PP vote shares have varied in a range of 10 percentage points, with an inter-quartile range of 4 percentage points. Morgan has the next largest bias in these data, resulting in an average underestimate of Coalition support of 4.9 percentage points; note that this estimate is based on treating the Morgan house effect as constant across the face-to-face polls and the one, final phone poll it fielded. Finally, Newspoll is estimated to have a 2.7 percentage point bias towards Labor that is statistically significant at the conventional 95% level. Recall that the pre-election Newspoll estimated Coalition vote share at 50%, based on a large sample of 2500 respondents; the disparity between this poll and the actual election result almost exactly mirrors the bias estimate, as it does for the other polling organisations. Newspoll fares better in estimating first preferences, with the bias parameter indistinguishable from zero at the conventional level of statistical significance. Both the ANU/ninemsn Internet poll and Morgan have large and statistically significant biases in their estimates of Coalition first preference vote shares, again systematically underestimating Coalition support.

Galaxy and Nielsen are estimated to have negligible biases in both their estimates of 2PP and first preference vote intentions (in the sense that the corresponding bias parameters cannot be distinguished from zero at conventional levels of statistical significance). Again, these negligible bias estimates reflect the fact that Galaxy and Nielsen's final poll estimates were quite close to the actual election outcome. Although Nielsen's last poll estimated the Coalition's 2PP vote share at 54% (an overestimate), over the course of the campaign Nielsen (like most polling organisations) tended to underestimate Coalition support, resulting in a negative bias estimate, albeit not statistically significant at conventional levels. Galaxy appears to have had an excellent campaign, with its estimated bias parameters almost exactly zero for both first preferences (0.4 percentage points, plus or minus about 2 percentage points) and 2PP (0.1 percentage points, again plus or minus about 2 percentage points).

It is also telling that the average bias of Nielsen, Galaxy and Newspoll is virtually zero on first preferences, but the 2PP average bias is over a percentage point

(underestimating Coalition 2PP support) and on the cusp of statistical significance. Differences in the ways the polling houses arrive at 2PP estimates are certainly implicated here: eg both Morgan and Newspoll have larger 2PP bias estimates than they do on first preferences, while Galaxy used minor-party preference flows observed at the 2001 election to estimate 2PP; see Brent (2004) for additional detail.

The house effects estimated here arise from numerous sources (as discussed above). But it is interesting to note that the largest house effects are associated with the mode of interview. The ANU/ninemsn Internet poll produces the largest biases, followed by Morgan (face to face), and the three telephone polls. The probability that the ANU/ninemsn Internet poll has a larger bias than the Morgan poll is 0.92 (in estimating 2PP vote shares), while the probability that the Morgan poll has a larger bias than the telephone polls is greater than 0.99. The average bias for the phone-based polling organisations (ACNielsen, Galaxy, Newspoll) is small—just over a percentage point underestimate of the Coalition's 2PP vote share—and on the threshold of conventional levels of statistical significance;[8] these three phone-based polls collectively did quite well in estimating first preference vote share, with the average bias almost exactly zero (0.1 of a percentage point, plus or minus about 1.8 percentage points).

Although the magnitudes of the biases do map onto mode of interview rather neatly, there is much more to house effects than interview mode. For instance, it may be that sampling and weighting procedures rather than mode underlie the poor performance of the self-selected Internet poll (ANU/ninemsn) and face-to-face polls (all but the final Morgan poll). This distinction seems especially important when considering the ANU/ninemsn poll, where mode (self-completion via the Internet) is confounded with sampling (self-selection). Although the ANU/ninemsn poll has the largest bias of the five polls I consider here, it would seem hasty to write off the Internet as a polling medium: Internet polls offer the promise of precision via large sample sizes (note that the marginal cost of an additional Internet respondent is close to zero) and timeliness, and may well be the wave of the future (eg McAllister 2004), provided the thorny issue of the non-representativeness of a self-selected sample of Internet users can be solved. Clearly, the large bias estimate attaching to the ANU/ninemsn Internet poll suggests that considerable work needs to be done on this score. The problem confronting the use of the Internet as a tool for election polling in Australia, at least in the near term, is that (1) compulsory voting means that the relevant population is the entire adult citizenry; (2) characteristics which predict Internet usage in Australia appear to be correlated with politically relevant characteristics, in particular age and the urban/rural divide (see, for example, Australian Bureau of Statistics 2003; Reid 2002). Deriving weighting schemes so as to make self-selected Internet samples representative of broadly defined populations is a topic of considerable interest in both academic and commercial circles (eg Berrens et al 2003; Schonlau et al 2004), and the future does look promising.

[8] We would reject the null hypothesis that the average bias of the three phone polls is zero, in favour of the one-sided alternative at conventional 95% levels of statistical significance; the probability that the average bias of the phone polls leads to an underestimate of Coalition 2PP is 0.93, just short of the conventional 95% standard.

Conclusion

I have presented a statistical model for dealing with two problems with election polls. By pooling and smoothing the polls, we obtain a more precise estimate of underlying vote intentions than can be formed from any single poll. By constraining the estimated trajectory of daily vote intentions to culminate with the known election result, it is possible to estimate bias parameters for each polling organisation, and, in turn, to recover day-by-day estimates of vote intentions that are purged of the biases afflicting any one polling organisation.

The estimates of daily vote shares suggest that there was a 2 percentage point shift towards the government over the course of the campaign, almost all of it coming in the opening weeks of the campaign. Substantial biases specific to polling companies are also apparent: Morgan and the ANU/ninemsn Internet polls systematically underestimate Coalition support by large magnitudes. On the other hand, the average bias of the telephone-based polls is small, and on the threshold of statistical significance at conventional levels.

The analysis reported here exploits the luxury of hindsight. The fact that the election outcome is known plays a powerful role in my analysis; the known election outcome provides a fixed point from which the daily stream of 2PP estimates and house effects are anchored. This is not to say that the model I present here is of no utility to researchers tracking shifts in vote support over the course of a campaign. Indeed, the model presented here can be easily deployed in 'real time', but the researcher needs to first make a commitment on the question of house effects. One strategy might be to ignore the house effects, naively pooling the polls with the bias parameters all (implicitly) set to zero; the results presented here strongly suggest that the house-specific biases are not ignorable, certainly when we admit Morgan and the ANU/ninemsn Internet poll to the pool. Ignoring house effects might work *if* one focused solely on phone polls *and* the house effects estimated here are in fact more or less fixed features of the various polling organisations. A better strategy, to be implemented in the next federal election campaign, is to use the house-specific bias estimates reported here to calibrate and then pool the estimates produced by each polling organisation.

References

Australian Bureau of Statistics. 2003. 'Household Use of Computers and the Internet.' *Australia Now* series. < http://www.abs.gov.au/ausstats/abs@.nsf/46d1bc47ac9d0c7bca256c470025ff8%7/c1d866341d03d9e9ca256d39001bc362!OpenDocument>.

Baum, M.A. and S. Kernell. 2001. 'Economic Class and Popular Support for Franklin Roosevelt in War and Peace.' *Public Opinion Quarterly* 65: 198–229.

Beck, N. 1990. 'Estimating Dynamic Models Using Kalman Filtering.' In *Political Analysis*, ed. James A. Stimson, vol. 1. Ann Arbor: University of Michigan Press: 121–56.

Berrens, R.P., A.K. Bohara, H. Jenkins-Smith, C. Silva and D.L. Weimer. 2003. 'The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples.' *Political Analysis* 11: 1–22.

Brent, P. 2004. 'For Whom the Polls Tell.' *Walkley Magazine* 30: 17.

Carlin, B.P., N.G. Polson and D.S. Stoffer. 1992. 'A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modelling.' *Journal of the American Statistical Association* 87: 493–500.

Carter, C.K. and R. Kohn. 1994. 'On Gibbs Sampling for State Space Models.' *Biometrika* 81: 541–53.

Erikson, R., C. Panagopoulos and C. Wlezien. 2004. 'Likely (and Unlikely) Voters and the Assessment of Campaign Dynamics.' *Public Opinion Quarterly* 68: 588.

Fleiss, J.L., B. Levin and M.C. Paik. 2003. *Statistical Methods for Rates and Proportions*, 3rd ed. Chichester: Wiley.

Goot, M. 2000. 'The Performance of the Polls.' In *Howard's Agenda*, eds Marian Simms and John Warhurst. St Lucia: University of Queensland Press: 37–47.
Goot, M. 2002. 'Turning Points: For Whom the Polls Told.' In *2001: The Centenary Election*, eds John Warhurst and Marian Simms. St Lucia: University of Queensland Press: 63–92.
Goot, M. 2005. 'The Polls: Liberal, Labor, or Too Close to Call.' In *Mortgage Nation*, eds Marian Simms and John Warhurst. Perth: API Network/Edith Cowan University Press.
Gow, D. 2001. 'Sampling in the Australian Election Studies: Past Practices and Future Prospects.' Presented to the Australian Election Study Workshop, Australian National University, Canberra.
Green, D.P., A. Gerber and S.L. De Boef. 1999. 'Tracking Opinion Over Time: A Method for Reducing Sampling Error.' *Public Opinion Quarterly* 63: 178–92.
Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.
Harvey, A.C. 1989. *Forecasting, Structural Time Series Models and the Kalman Filter*. New York: Cambridge University Press.
Jackman, S. 2000a. 'Estimation and Inference are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation.' *Political Analysis* 8: 307–32.
Jackman, S. 2000b. 'Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo.' *American Journal of Political Science* 44: 375–404.
McAllister, I. 2004. 'A Margin for Error.' *The Bulletin* 19 October.
McDermott, M. and K.A. Frankovic. 2003. 'Horserace Polling and Survey Method Effects: An Analysis of the 2000 Campaign.' *Public Opinion Quarterly* 67: 244–64.
Reid, A. 2002. *Town and Country and the Digital Divide*. Sydney: ACNielsen/NetRatings.
Schonlau, M., K. Zapert, L.P. Simon, K.H. Sanstad, S.M. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner and S.H. Berry. 2004. 'A Comparison between Responses from a Propensity-weighted Web Survey and an Identical RDD Survey.' *Social Science Computer Review* 22: 128–38.
Stimson, J.A. 1991. *Public Opinion in America: Moods, Cycles, and Swings*. Boulder: Westview.
West, M. and J. Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. New York: Springer.
Wolfers, J. and A. Leigh. 2002. 'Three Tools for Forecasting Federal Elections: Lessons from 2001.' *Australian Journal of Political Science* 37: 223–40.

Appendix: Bayesian Estimation and Inference via Markov Chain Monte Carlo

Equations (3), (4) and (5) define the statistical model used for pooling and smoothing the polls, as well as estimating house effects. The unknown parameters are (1) α_t, the Coalition's vote share on day t, $t = 1, \ldots, T - 1$; (2) δ_{j_i}, the house effect of polling organisation j_i, where j indexes the set of five polling organisations analysed here; and (3) ω^2, a variance parameter tapping the magnitude of day-to-day variability in the α_t.

Inference for these parameters is via Bayesian methods, meaning that we require a characterisation of the posterior density of the model parameters, 'posterior' literally in the sense of after having observed the polling data. I rely on computationally intensive, simulation-based methods to provide this characterisation; specifically, I use Markov chain Monte Carlo (MCMC) techniques, and in particular Gibbs sampling, to provide a large number of approximately independent samples from the posterior density of the model parameters. In this case the Gibbs sampler was run for 26 million iterations, discarding the first one million iterations as burn-in, and each 1000th iteration of the remaining 25 million iterations retained for inference, yielding 25,000 approximately independent samples from the posterior density of the model parameters. Point estimates for parameters reported in the body of the paper are the average of the sampled values for the respective parameters; 95% confidence intervals are formed by computing the 2.5 and 97.5 percentiles of the sampled values. Reviews of MCMC methods tailored for a political science audience can be found in Jackman (2000a, b). I adopt the Bayesian approach to dynamic, latent-variable modelling as discussed in West and Harrison (1997), Carter and Kohn (1994) and Carlin, Polson and Stoffer (1992). Complete details on the implementation of the

Gibbs sampler for this problem, and the prior distributions assumed, are available in the longer, working paper version of this article on the author's Website.

Prior distributions are a critical component of Bayesian modelling, a formal statement of the researcher's a priori beliefs about the model parameters. Equations (5) and (6) supply priors for the α_t parameters. For the house-effects parameters δ_j I use a vague normal prior centred at zero $\delta_j \sim N(0, d^2)$, where j indexes survey organisations, and with d an arbitrary large constant; I use $d^2 = (0.15/2)^2 = 0.005625$.

A prior distribution is also required for the variance of the day-to-day changes in voter sentiment, ω. I use a uniform prior:

$$f(\omega) = \begin{cases} 100 & \text{if } 0 < \omega < 0.01 \\ 0 & \text{otherwise} \end{cases} \tag{A1}$$

That is, I presume that day-to-day changes are not massive; even at this maximum prior value of $\omega = 0.01$ (1 percentage point), 95% of the daily changes in the α_t are no larger than plus or minus 2 percentage points.