ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)

An assessment of the causes of the errors in the 2015 UK general election opinion polls

Patrick Sturgis,

University of Southampton, UK

Jouni Kuha,

London School of Economics and Political Science, UK

Nick Baker,

Quadrangle, London, UK

Mario Callegaro,

Google, London, UK

Stephen Fisher,

University of Oxford, UK

Jane Green,

University of Manchester, UK

Will Jennings,

University of Southampton, UK

Benjamin E. Lauderdale

London School of Economics and Political Science, UK

and Patten Smith

Ipsos-MORI, London, UK

[Received February 2017. Revised August 2017]

Summary. The opinion polls that were undertaken before the 2015 UK general election underestimated the Conservative lead over Labour by an average of 7 percentage points. This collective failure led politicians and commentators to question the validity and utility of political polling and raised concerns regarding a broader public loss of confidence in survey research. We assess the likely causes of the 2015 polling errors. We begin by setting out a formal account of the statistical methodology and assumptions that are required for valid estimation of party vote shares by using quota sampling. We then describe the current approach of polling organizations for estimating sampling variability and suggest a new method based on bootstrap

Address for correspondence: Patrick Sturgis, Department of Social Statistics and Demography, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: P.Sturgis@soton.ac.uk

© 2017 The Authors Journal of the Royal Statistical Society: Series A (Statistics in Society) 0964–1998/17/181000 published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

resampling. Next, we use poll microdata to assess the plausibility of different explanations of the polling errors. Our conclusion is that the primary cause of the polling errors in 2015 was unrepresentative sampling.

Keywords: Election polling; Late swing; Quota sampling; Turnout weighting; Unrepresentative samples

1. Introduction

The result of the 2015 UK general election came as a shock to most observers. During the months and weeks leading up to election day on May 7th, the opinion polls consistently indicated that the outcome was too close to call and the prospect of a hung Parliament therefore appeared almost inevitable. Although there was some variation across polling companies in their estimates of the party vote shares, their estimates of the difference between the Conservative and Labour Parties exceeded 2 percentage points in only 19 out of 91 polls during the short campaign from March 30th, with 0 as the modal estimate of the Conservative lead.

The poll-induced expectation of a dead heat undoubtedly informed party strategies and media coverage during the campaign and may ultimately have influenced the result itself, albeit in ways that are difficult to determine satisfactorily. In the event, the Conservative Party won a narrow parliamentary majority, taking 37.7% of the popular vote in Great Britain (and 330 of the 650 seats in the House of Commons), compared with 31.2% for the Labour Party (232 seats; see Hawkins *et al*. (2015) for the official results). The magnitude of the errors on the Conservative lead, as well as the consistency of the error across polling companies (henceforth referred to as 'pollsters') strongly suggests that systematic factors, rather than sampling variability, were the primary causes of the discrepancy.

Table 1 presents the final published vote intention estimates for the nine pollsters that were members of the British Polling Council (BPC) at the time of the election, plus three non-members who published estimates. These are estimates for Great Britain excluding Northern Ireland, which is the usual population of inference for election polls in the UK. The estimates for the smaller parties are close to the election result, with mean absolute errors of 0.9%, 1.4%, 1.3% and 1.1% for the Liberal Democrats, UK Independence Party, the Green Party and other parties (combined) respectively, all of which are within the pollsters' notional margins of error for party shares due to sampling variability (which are usually stated as $\pm 3\%$ for point estimates). However, for the crucial estimate of the difference between the two main parties, 11 of the 12 Great Britain polls in Table 1 were some way from the true value, and attention has naturally focused on this error. Whereas the election result saw Labour trail the Conservatives by 6.5 percentage points, five polls in the final week reported a dead heat, three reported a 1% lead for the Conservatives, two a 1% lead for Labour and one a 2% lead for Labour. For all nine BPC members, the notional $\pm 3\%$ margin of error does not contain the true election result. SurveyMonkey published the only final poll to estimate the lead correctly, although their estimates were too low for both the Conservatives and Labour and, indeed, had higher mean absolute errors across all parties than the average of the other polls.

In Scotland, the three polls that were conducted in the final week overestimated the Labour vote share by an average of 2.4 and underestimated the Scottish National Party share by 2.7 percentage points. The average error of 5.1 percentage points on the lead of the Scottish National Party over Labour in Scotland was only slightly smaller than the average error on the lead of the Conservatives over Labour in the polls for Great Britain.

These errors were not just a cause of embarrassment for the pollsters. Media sponsors publicly questioned the quality and value of the research that they had commissioned, with at least one national newspaper stating that it would afford less prominence to election polling in its political

Table 1. Published estimates of voting intention for various parties (as the percentage of vote in Great Britain), from the final polls before the UK general election on May 7th, 2015

Pollster	*Survey mode*	*Days of fieldwork*	*Sample size*	*Results for the following parties (%):*					
				Conservative	*Labour*	*Liberal Democrats*	*UK Independence Party*	*Green*	*Other*
Populus	On line	May 5th–6th	3917	34	34	9	13	5	6
Ipsos-MORI	Phone	May 5th–6th	1186	36	35	8	11	5	5
YouGov	On line	May 4th–6th	10307	34	34	10	12	4	6
ComRes	Phone	May 5th–6th	1007	35	34	9	12	4	6
Survation	On line	May 4th–6th	4088	33	34	9	16	4	4
ICM	Phone	May 3rd–6th	2023	34	35	9	11	4	7
Panelbase	On line	May 1st–6th	3019	31	33	8	16	5	7
Opinium	On line	May 4th–5th	2960	35	34	8	12	6	5
TNS UK	On line	April 30th–May 4th	1185	33	32	8	14	6	6
Ashcroft†	Phone	May 5th–6th	3028	33	33	10	11	6	8
BMG†	On line	May 3rd–5th	1009	34	34	10	12	4	6
Survey Monkey†	On line	April 30th–May 6th	18131	34	28	7	13	8	9
Election result				37.7	31.2	8.1	12.9	3.8	6.4
Mean absolute error				3.9	2.7	0.9	1.4	1.3	1.1

†Not members of the BPC in May 2015.

coverage in the future. Politicians and peers suggested that the polling inaccuracies had affected the outcome of the election, speculating that Labour might have done better if the polls had been accurate. A private members' bill was introduced in the House of Lords on May 28th, 2015, proposing state regulation of the polling industry (*Regulation of Political Opinion Polling Bill [HL] 2015-16*). Concern was also expressed by social and market research industry professionals; as the most direct way that the public encounters survey and opinion research, it was feared that the failure of the polls might have negative consequences for public confidence in social and market research and official statistics more generally.

It is therefore important that we understand what went wrong with the general election opinion polls in 2015, so that the risks of similar failures in the future are reduced. This is our objective in this paper. Similar investigations have been carried out in the aftermath of previous historical polling errors, both in the UK (Market Research Society, 1994) and in the USA (Crossley, 1936; Mosteller *et al.*, 1949; Traugott *et al.*, 2009; American Association for Public Opinion Research, 2017) and have resulted in important changes to the conduct and reporting of polls (Converse, 1987).

We draw here on the findings and conclusions that were set out in the report of the inquiry into the failure of the polls in 2015 that was established by the BPC and Market Research Society (Sturgis *et al.*, 2016). In addition to the material that is contained in that report, we provide a more detailed and formal account of the methodology of vote share estimation by using opinion polls, drawing out the key assumptions on which the methodology is based and using this to structure our presentation and interpretation of findings. We also set out a new procedure which can be used to produce estimates of the sampling variability of opinion polls collected by using

quota sampling, which better reflects their design than the (sample size invariant) ±3% rule of thumb for the 'margin of error'.

The remainder of the paper is structured as follows. In Section 2 we describe the methodology of the 2015 opinion polls, the assumptions required for valid point estimation and the new methodology that we propose for variance estimation. The data that we used to evaluate the causes of the polling errors is described in Section 3 and the results and interpretation of our analyses are in Section 4, where we focus on the three key potential factors: late swing, turnout weighting and sampling. Our conclusion from these analyses is that the polling miss in 2015 occurred primarily because the procedures that were used by the pollsters to recruit respondents produced samples which were unrepresentative of the target population's voting intentions. These biases were not mitigated by the statistical adjustments that pollsters applied to the raw data. Other factors made, at most, a very modest contribution. Concluding remarks are given in Section 5.

Data and code which are illustrative of the kinds of data that are analysed in the paper and the program that was used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. The methodology of pre-election polls

2.1. Point estimation of vote shares

The polls that were conducted before the 2015 general election employed one of two data collection modes: on-line self-completion or computer-assisted telephone interviewing. For the selection of respondents, all the polls employed non-probability (quota) rather than probability sampling, so for estimation they cannot directly apply the well-established theory for probability sampling (Kish, 1965; Groves *et al*., 2009). The operational procedures that were employed to recruit respondents were diverse and incorporated a range of random and purposive selection mechanisms (see Sturgis *et al.* (2016) for a more detailed account of these procedures). All British pollsters, however, took a common general approach to sampling and estimation: they assembled a quota sample of eligible individuals, calculated a weight to match the sample to known population distributions for a set of auxiliary variables and a weight to account for differential likelihood of voting. They then combined these two weights and produced weighted estimates of vote intention for the population of voters from the sample data.

It is useful for our later evaluation of the potential causes of the polling errors to describe this general approach in more formal terms. Our specification here draws on previous treatments of the assumptions that are required for the validity of point estimation by using quota sampling (Smith, 1983; Deville, 1991), extended to accommodate the inclusion of turnout probabilities. It is important to note that we do not claim that this is how the pollsters explicitly motivate their methodology. It is, nonetheless, implicit in the procedures as they are implemented.

We first define a set of variables which are relevant for the estimation of party vote shares for the target population. These are all characteristics of individuals and are, in practice, treated as categorical variables, whatever their natural metric. We denote by X auxiliary variables which will be used to derive weights to match population distributions, and by L additional variables which will be used to predict the probability that an individual will vote in the election. In a typical poll, X includes characteristics such as sex, age, region and social class, as well as measures of party identification or vote in a previous election, whereas L is an individual's self-assessment of how likely he or she is to vote in the election. Further, let V denote the party that the individual reports he or she intends to vote for (after 'Don't know' answers and refusals have

been dropped or imputed to specific parties), T an indicator of whether the individual actually voted in the election (with $T = 1$ for yes and $T = 0$ for no), P the party (if any) that they actually voted for and S an indicator of whether or not an individual is included in the sample ($S = 1$ for yes and $S = 0$ for no).

The target population of the poll should be such that it includes all individuals who are eligible and registered to vote in the election (but it can also include people who are not, assuming that they will be filtered out later by being assigned turnout probabilities of 0), and that distributions of weighting variables in the population are known. In the polls that are considered here, this population is typically that of adult residents of Great Britain (even though this has the shortcoming that it omits voters who live abroad). Consider X partitioned as $(X_{(1)}, \ldots, X_{(p)})$, where the subsets $X_{(j)}$ are such that their distributions $p(X_{(j)})$ in the population are assumed known from the census or other sources (we denote marginal and conditional distributions of variables by $p(\cdot)$ and $p(\cdot|\cdot)$). The $X_{(j)}$ are typically univariate, although with some exceptions (e.g. age distribution may be specified separately by sex).

When interviews have been completed, weights are created in such a way that the weighted distributions of all $X_{(j)}$ in the sample match their population distributions $p(X_{(j)})$. This step is similar to calibration weighting of probability samples (Deville and Särndal, 1992), so we refer to the resulting weights as *calibration weights*. In the election polls that are considered below all the weighting variables $X_{(j)}$ were categorical, in which case calibration is equivalent to raking and it can be carried out by using, for example, the iterative proportional fitting algorithm (Deming and Stephan, 1940).

By way of illustration, suppose that X consists of age by sex, $X_{(1)}$, region of residence, $X_{(2)}$, and vote in the most recent previous election, $X_{(3)}$, and that $X_{(1)}$ and $X_{(2)}$ are used also as quota variables, with $p(X_{(1)})$ and $p(X_{(2)})$ as their target distributions. In quota sampling, the aim is then to obtain a sample where the distribution of age by sex and the marginal distribution of region of the sample match the target distributions, at least to a close approximation (an exact match is often not achieved in practice). Next, a raking algorithm is applied to this sample, with X as the weighting variables. The resulting calibration weights will be such that the weighted distributions of age by sex, region and past vote match their population distributions $p(X_{(1)})$, $p(X_{(2)})$ and $p(X_{(3)})$ exactly.

The goal of a vote intention poll is to estimate $p(V|T = 1)$ in the population, i.e. the distribution of responses to the question on party choice among those members of the population who will turn out to vote. This can be expressed as

$$p(V|T=1) = \frac{\sum_{X,L} p(T=1|V,L,X)\, p(V,L|X)\, p(X)}{p(T=1)} \tag{1}$$

where the sum is over the possible values of L and X. Here $p(X)$, $p(V, L|X)$ and $p(T = 1|V, L, X)$ describe the population distribution of the weighting variables, distribution of voting intention and stated likelihood to vote, and probability of turnout respectively. To estimate this quantity, a poll draws a sample of respondents ($S = 1$), selected through quota sampling with quota targets defined by a subset of X and elicits values of (X, L, V) from the sampled respondents via questionnaire. Turnout T is not known at the time of the poll, except for respondents who have already voted by post. Letting $i = 1, \ldots, n$ index the sampled respondents, the calibration weights w_i^* are then calculated. The distribution of (V_i, L_i, X_i) in the sample, with weights w_i^*, is used as an estimate of $p(V, L, X) = p(V, L|X)p(X)$ in the population. Next, let p_{T_i} denote values of $p(T_i = 1|V_i, L_i, X_i)$ assigned for each respondent from an assumed model for the turnout probabilities, and define $w_i = p_{T_i} w_i^*$. Letting $I(V_i = v)$ be an indicator variable for any particular

party v which is 1 if a respondent's stated vote intention is $V_i = v$ and 0 otherwise, the vote intention proportions for the parties are estimated by the weighted proportions

$$\hat{p}(V = v|T = 1) = \frac{\sum_{i=1}^{n} I(V_i = v) w_i}{\sum_{i=1}^{n} w_i}. \quad (2)$$

Using equation (2) to estimate $p(V = v|T = 1)$ implies a number of assumptions about the quantities on the right-hand side of equation (1). First, it is assumed that the p_{T_i} assigned to respondents are equal to the probabilities $p(T_i = 1|V_i, L_i, X_i)$ under the conditional distribution of turnout given (V, L, X) in the population. Second, it is assumed that $p(V, L|X, S = 1) = p(V, L|X)$, i.e. that the (V_i, L_i) in the sample (unweighted, since the weights w_i^* are constant given X) can be treated as random variables drawn from their distribution in the population, at each level of the variables X which are used to derive the calibration weights. We refer to this as the assumption of representative sampling. It is weaker than the requirement of representativeness given only the quota variables, which are typically only a subset of X.

These two assumptions are still not sufficient for valid estimation of $p(V = 1|T = 1)$ because the calibration weights ensure only that the weighted distributions in the sample match the population distributions for the marginal distributions of $X_{(j)}$ but not for the joint distribution $p(X)$ in equation (1). This weighted joint distribution of X in the sample matches the full $p(X)$ only if the sample is (fortuitously) representative in the higher order associations among the X which have not been fixed to match population totals. Alternatively, estimation with equation (2) is also valid if the true conditional distributions of (V, L) and T are such that only the $p(X_{(j)})$ actually contribute to probability (1). This is so, for example, if both $p(T = 1|V, L, X)$ and $p(V, L|X)$ are linear functions of their explanatory variables and the product of these functions does not involve any products of $X_{(j)}$ and $X_{(k)}$ $(j \neq k)$. This is true, for instance, in cases where $p(V, L|X)$ does not depend on interactions between the $X_{(j)}$ and $p(T = 1|V, L, X) = p(T = 1|V, L)$ does not depend on X.

If these assumptions hold, it is possible to estimate the distribution of stated vote intentions V among eventual voters. What commissioners and consumers of polls really want to know, however, is not the distribution of V but the distribution of actual votes in the election, which we denote by P. A pre-election poll cannot, however, provide direct information about P because P does not exist (except for postal voters) until election day. To interpret the poll estimates by using equation (2) as actual vote shares, it must also be assumed that $p(V|T = 1) = p(P|T = 1)$. This will be true if $V_i = P_i$ for every individual, but also if individual level changes between V_i and P_i are self-cancelling in the aggregate.

In summary, the key assumptions which underlie the estimates of pre-election polls as they were conducted for the 2015 UK general election are as follows.

Assumption 1 (representative sampling). Given any value of the weighting variables X, observations (V_i, L_i) in the poll can be treated as a random sample (with equal inclusion probabilities) from $p(V, L|X)$ in the population.

Assumption 2 (correct model for turnout probabilities). The assigned turnout weights p_{T_i} are equal to the probabilities $p(T_i = 1|V_i, L_i, X_i)$ from the conditional distribution of T which holds in the population.

Assumption 3 (agreement between stated vote intention and actual vote). $p(V|T = 1) = p(P|T = 1)$, i.e. there is no difference between the stated vote intention in the final poll and the choice that is made in the election, or where such discrepancies do exist they are self-cancelling in the aggregate.

These are made together with the additional conditions on the distributions of X, (V, L) and T that were discussed above. If assumptions 1 and 2 hold, equation (2) provides consistent estimates of the vote intentions $p(V|T=1)$ and, if assumption 3 holds as well, of the actual vote shares $p(P|T=1)$. It is unlikely in practice that these assumptions will be exactly satisfied, so it is better to regard them as ideal conditions that the polls should aim to be as close to as possible to produce reasonable estimates.

These assumptions are stringent. Assumption 1 requires that the samples are representative given the weighting variables, even though in quota sampling the sampling probabilities are not known and will probably be 0 for some members of the population, and robust population data for weighting is limited. Assumption 1 will be violated if, over hypothetical repetitions of the sampling process, the distribution of vote intention, given the weighting variables, is different from the corresponding population distribution. For instance, in the example that we introduced earlier this would occur if, for some combinations of the weighting variables age-by-sex, region and past vote, the respondents are more educated than the target population, and education is associated with vote even given the weighting variables. Assumption 2 requires that turnout probabilities can be modelled to a high degree of accuracy, even though there is little evidence on which to base such a prediction. Assumption 3 requires that respondents' reported pre-election vote intentions accurately represent their actual votes, so this can fail if there are missing data or measurement error in the stated vote intention, or if reported vote intentions change between the poll and the election. All of these conditions are prone to violation at any given election and may fail in ways that cause large errors in estimated vote shares. In Section 4 we examine evidence of such failures in the 2015 polls for each of the assumptions.

2.2. Sampling variability of point estimates

We concluded from Table 1 that the 2015 polling miss was not due to random sampling error. However, this conclusion is based on the rather unsatisfactory notion of a $\pm 3\%$ 'margin of error' applied to any point estimate for a proportion, which is currently used by UK pollsters. This rule of thumb is derived under an 'as if' assumption of simple random sampling for a sample of size of 1000, which is a common sample size for opinion polls. This heuristic is clearly not appropriate for the sample designs of the 2015 polls. Yet, ignoring their sampling variability is clearly unsatisfactory and, indeed, the recent American Association for Public Opinion Research task force on non-probability sampling recommended that '... users of non-probability samples should be encouraged to report measures of the precision of their estimates' (Baker *et al.*, 2013). Here, we propose a method of calculating the precision of poll estimates from quota samples, which better reflects the nature of the sampling and estimation procedures in the election polls.

We propose a bootstrap resampling method which involves the following three steps:

(a) draw M independent samples by sampling respondents from the full achieved sample, with replacement and in a way which matches the quota sampling design;
(b) for each sample thus drawn, calculate the point estimates of interest in the same way as for the original sample, including calibration and turnout weighting;
(c) use the distribution of the estimates from the M resamples to quantify the uncertainty in the poll estimates.

This draws on the basic ideas of bootstrap estimation in general (Davison and Hinkley, 1997) and for probability samples in particular (Wolter, 2007). It is worth emphasizing that a key part of the procedure is that the calibration weights are calculated afresh for each bootstrap sample in step (b) (Rust and Rao, 1996). For non-probability samples, a comparable approach

has been proposed by de Munnik *et al.* (2013), although they used it to assess the quality of a sampling design by resampling from a simulated population, rather than the sample itself. An alternative approach to estimating uncertainty would be to adapt variance formulae that are used with probability sampling under approximate assumptions about the nature of the quota samples (Deville, 1991; Rivers, 2013). It would be more difficult using this approach, however, to accommodate specific features of the poll estimation, such as calibration and turnout weighting, which are easily accounted for in the bootstrap method.

The intuitive idea of the bootstrap method is that the samples from step (a) are used to represent the variation from one sample to the next that would be observed if the quota sampling procedure were implemented repeatedly. Applying the estimation procedure to each of the bootstrap samples then represents the sample-to-sample variation in the resulting estimates. It is important to note that bootstrapping thus produces an estimate of sampling variance and not of the variability of estimates around their true values (i.e. mean-squared errors), unless the assumptions that were stated in Section 2.1 are satisfied and the estimates are thus approximately unbiased.

A challenge for the resampling step (a) is that the sampling quotas do not define a partitioning of the population, so the resample cannot be implemented within the quota as if they were sampling strata. Instead, the quota form a set of separate targets for different variables, all of which should have been reached when the sampling has been completed. To implement the resampling in this situation, the analyst would ideally know the exact procedures through which the quota sampling was implemented, but these specific details are not available to us. In our calculations for the 2015 polls, we have therefore used the following algorithm which represents the generic features of quota sampling. First, we set the quota targets to be the realized sample distributions of the quota variables that were used by a given pollster. In the first iteration of the resampling, the pool of potential respondents is the full observed sample, from which we draw a sample of the same size as the full sample, but with replacement. We then drop from this first iteration sample any observations which overfill a quota category, and we retain the rest. For the next iteration of the sampling, the pool of potential respondents now consists only of those who are in quota categories which remain to be filled. The sample size of the second iteration is now the number of observations that need to be added to reach the original sample size. In other words, at each iteration the retained sample is 'topped up' through a resample drawn from the quota categories which are not yet full. Additional iterations continue until all the quotas are full, or until there are no respondents in the original sample who belong to all the incomplete quota categories at once. In the latter case we could run the algorithm again, or use the sample that is obtained at this point, even though it is slightly smaller than the observed sample. For the estimates that are presented here, we used the latter strategy. A more detailed statement of the algorithm, computer code and an example are included in the on-line supplementary materials to this paper.

Results from the bootstrap estimation of sampling variability in the final polls are presented in Table 2, based on 10 000 bootstrap samples. It shows point estimates and 95% interval estimates of the Conservative–Labour difference in vote shares. These are adjusted percentile intervals (Davison and Hinkley, 1997), although standard percentile intervals and symmetric normal intervals give similar results. None of the intervals in Table 2 includes the election result of a 6.5-point Conservative lead. We can therefore be confident in our initial conclusion that the polling miss was not due to sampling variability.

Table 2 also shows bootstrap standard errors of the estimated vote shares for the Conservatives and for Labour. We can compare these with the notional 'margin of error' that is obtained if the poll sample is treated as a simple random sample, where the sampling variance for an estimated vote share p from a sample of size n is given by $p(100 - p)/n$. Table 2 shows estimated design

Table 2. Measures of uncertainty in estimates of voting intention from the final polls: point estimates and 95% interval estimates for the Conservative–Labour difference, and standard errors se and estimated design effects d^2 for the Conservative and Labour vote shares†

Pollster	*Survey mode*	*Conservative–Labour (%) (election result, 6.5%)*		*Conservative (%)*		*Labour (%)*		*n*
		Estimate	*95% interval‡*	*se*	d^2	*se*	d^2	
Populus	On line	−0.1	(−2.5; 2.0)	0.7	0.82	0.7	0.73	3695
Ipsos-MORI	Phone	−0.3	(−6.6; 6.1)	1.8	1.37	1.9	1.40	928
YouGov	On line	0.4	(−1.1; 1.8)	0.4	0.76	0.4	0.76	9064
ComRes	Phone	0.8	(−4.6; 6.3)	1.5	0.86	1.9	1.30	852
Survation	On line	0.1	(−2.2; 2.5)	0.7	0.79	0.7	0.85	3412
ICM	Phone	0.0	(−2.8; 3.1)	0.9	0.53	1.0	0.68	1681
Panelbase	On line	−2.7	(−5.6; 0.2)	0.9	1.17	0.9	1.16	3019
Opinium	On line	0.4	(−1.8; 2.5)	0.6	0.42	0.7	0.47	2498
TNS UK	On line	0.8	(−3.6; 5.2)	1.4	0.79	1.3	0.72	889

†The interval estimates and standard errors have been calculated from the microdata provided by the pollsters, using bootstrap resampling with 10000 bootstrap samples. Some of these replicated estimates differ slightly from the published results in Table 1, mainly because of rounding and differences in algorithms used for calibration weighting. n, number of respondents who gave a voting intention for a party. $d^2 = \mathrm{se}^2/\{p(1-p)/n\}$ where p is the estimated vote share.
‡Adjusted percentile interval.

effects d^2 for the vote shares, calculated by dividing the bootstrap variance by the variance under simple random sampling, with p as the estimated share and n the number of respondents who gave a vote intention (this n ignores the variability in the turnout probabilities and so probably underestimates the design effect). Most of the design effects are less than 1, indicating that the sampling variability is smaller than would be expected under simple random sampling. When this is so, the conventional margins of error somewhat overestimate the sampling variability in the poll estimates. The increased efficiency of the estimates is mainly due to calibration on variables measuring party affiliation or past voting, which are strongly correlated with current vote intention (all pollsters except Ipsos-MORI used this type of variable in their calibration weighting). If these variables are omitted from the quotas and weighting, all design effects in Table 2 are greater than 1, with values between 1.04 and 1.68.

3. Data

Our main evidence for assessing the source of the polling errors in 2015 is data from the polls themselves. Each of the nine BPC members provided respondent level microdata, together with documentation on their methodology, including fieldwork procedures, quota targets and weighting. These data were provided for the first, penultimate and final polls that were conducted during the short campaign but almost all of the analyses reported in this paper use the final polls (i.e. the nine polls that were conducted by the BPC members in Table 1). The six pollsters who carried out surveys where respondents were recontacted after the election also provided these data sets (these are discussed in Section 4.2). We could replicate all published estimates for these 27 pre-election polls, enabling us to exclude the possibility that flawed analysis or use of inaccurate weighting targets contributed to the polling miss.

We also analysed data from the 2015 rounds of the British Election Study (BES) and the British Social Attitudes (BSA) survey to benchmark the poll estimates against surveys which

use random-probability sample designs. The methodology of these surveys is described in detail elsewhere (Fieldhouse *et al.*, 2015; Clery *et al.*, 2016), but, in brief (and for both surveys), a multistage, stratified probability sample of addresses is drawn from the *Post Office Address File* and an interview is attempted with a randomly selected eligible adult at each eligible address. Multiple calls are made to each selected address at different times of day and on different days of the week to achieve an interview. Substitutions for sampled respondents who were not reached or who declined to be interviewed are not permitted. Interviews are carried out face to face by trained interviewers via questionnaires loaded onto laptop computers. The BES and BSA survey attained response rates of 56% and 51% (American Association for Public Opinion Research response rate 1) respectively, which, though not especially high in historical terms, are good by contemporary standards. The interviews were carried out after the election, in May–October 2015.

4. Assessment of potential causes of the polling errors

In this section we assess the evidence in support of the different potential causes of the errors in the 2015 polls. The discussion is structured around the three core assumptions that were set out in Section 2, with assumption 3 considered in Section 4.1, turnout weighting (assumption 2) in Section 4.2, and representative sampling (assumption 1) in Section 4.3.

In the polling inquiry report, the following more minor factors were also considered and dismissed as contributory causes of the polling errors: treatment of postal voters, overseas voters, voter registration, question wording and order, and mode of interview. Although we do not consider these factors directly in this paper, each can be understood as violations of the three key assumptions that are covered in the following sections. For instance, errors due to omission of overseas and postal voters would be violations of assumption 1 on representative sampling. Errors due to question wording or measurement mode would fall under violation of assumption 3, which stated that vote intention is equal to the actual vote. Likewise, unregistered voters would be a particular violation of assumption 2, that the turnout probabilities are correct. More detail on the specific reasons for ruling out these factors is provided in Sturgis *et al.* (2016). Nor do we discuss here the phenomenon of 'herding', which is when there is more consensus across poll estimates than would be expected under random sampling because it relates to the variability of estimates across polls, rather than to bias in poll estimates (Sturgis *et al.*, 2016).

4.1. Differences between reported vote intention and actual vote ('late swing')

Some voters agree to take part in opinion polls but do not disclose the party that they intend to vote for. Others do not know whom they will support, or they deliberately misreport their vote intention or report their intention truthfully but then change their minds after the poll. If a sufficient number of these types of voters move disproportionately to different parties between the polls and election day, vote shares that are estimated from the polls will differ from the election result. This discrepancy will not be due to inaccuracy of the polls as estimates of the stated vote intentions, but to inadequacy of the assumption (assumption 3) that the stated intentions V can be treated as a measure of the actual vote P. Following the convention in the polling literature, we refer to a difference between V and P as *late swing*. The term refers most naturally to a switching from one party to another, but we also include movement from non-substantive responses ('don't knows' and refusals) to a party choice.

Reports into the polling failures at the 1970 (Butler and Pinto-Duschinsky, 1971) and 1992 (Market Research Society, 1994) elections both attributed a prominent role to late swing. This

Table 3. Conservative lead over Labour (for Great Britain), estimated from five post-election recontact surveys and (for the same respondents, those who reported that they had voted) from polls before the election†

	Results for the following pollsters (%):					
	TNS	*Populus*	*ICM*	*Survation*	*YouGov*	*Combined*‡
Before election	−2.1	−1.3	0.0	0.3	0.5	−0.2
After election	1.9	−0.4	1.9	3.8	−0.8	0.4
Difference	4.0	0.9	1.8	3.4	−1.3	0.6
95% interval§	(−0.5; 8.5)	(−0.6; 2.4)	(−0.9; 4.6)	(0.2; 6.6)	(−2.3; −0.3)	(−0.2; 1.5)
Sample size	1477	3036	2480	1525	6712	15230

†The election result was a 6.5% Conservative lead.
‡Weighted average of the estimates for the pollsters, weighted by the sample sizes.
§Symmetric 95% confidence intervals, using standard errors estimated from 10000 bootstrap resamples for each pollster.

was particularly so for the 1970 report, which concluded that late swing was almost entirely to blame for the failure to predict the Conservative victory in that election. It has also been identified as a contributory factor for polling misses in the USA (American Association for Public Opinion Research, 2009; Keeter *et al*., 2016). There are, therefore, good *prima facie* grounds for assuming that late swing may have contributed to the polling miss in 2015.

The most direct way of assessing late swing is to examine data from *recontact surveys*, where the same respondents have been interviewed both before and after the election. Six pollsters carried out such surveys, although one proved to be unusable for our purposes because fieldwork outcomes did not distinguish between refusals and non-voters. For the analysis of late swing we use only the samples of voters who reported after the election that they had voted, which means that turnout weights are not needed and that assumption 2 is not required. Because not all respondents that were recontacted provided an interview, the estimates are weighted by the product of the pre-election calibration weight and an attrition weight. The attrition weight was calculated as the inverse of the predicted probability of responding to the recontact survey, derived from a logistic regression model where the predictor variables were all the variables that were used for weighting in the final poll, plus the question on likelihood to vote if used for the poll. For two pollsters the sample sizes for the recontact surveys were very small but in these cases it was possible to include respondents who were interviewed in earlier polls for the same company during the short campaign.

Table 3 shows point estimates of the Conservative lead over Labour for these five samples, from the pre-election polls and the recontact surveys. In four of the five polls the post-election estimates move in the direction of a larger lead for the Conservatives, and in one poll (the poll with the largest sample size) in the opposite direction. The average change towards the Conservatives weighting all five polls equally is 1.8 points and −0.4 points if only respondents from the final polls are included. An average of the five estimates (for all respondents) weighted by sample size is 0.6 points, with a 95% confidence interval (obtained by using bootstrap estimates of the standard errors of the estimates) is from −0.2% to 1.6%. Regardless of which of these estimates one prefers, none of them is nearly enough to explain the total error in the polls.

Some part of these changes may be due to measurement effects in the pre-election responses such as the treatment of refusals and ‘don’t know’ responses. Pollsters used different combina-

tions of 'squeeze' questions (where initially non-disclosing respondents are pushed to provide a vote intention) and imputation to allocate vote intentions for these respondents. However, the aggregate effects of these procedures on the final vote intention estimates were small (Sturgis *et al.*, 2016).

A frequently advanced explanation of polling errors that is by media commentators is deliberate misreporting, which is when respondents knowingly tell pollsters that they will vote for a particular party when they actually intend to vote for a different party. This is generally considered to occur, not out of capriciousness or spite against pollsters, but because of processes of social desirability. For example, in the UK, deliberate misreporting has been invoked to explain the tendency of polls to underestimate the Conservative vote as a result of respondents being unwilling to admit to voting Conservative—so-called 'shy Tories'. But the same phenomenon could apply to any party that voters feel embarrassed to admit to supporting because there is some sort of social stigma associated with it.

A response pattern of deliberate misreporting of voting intention is indistinguishable from late swing; the individual tells the pollster that they will vote for party A but subsequently votes for party B. Whether their initial report was a deliberate misreport is neither here nor there with regard to the pattern of response that is observed. Our conclusions about late swing therefore also enable us to rule out deliberate misreporting as a cause of the polling miss. A limitation to this conclusion is that actual vote could also be deliberately misreported in the recontact surveys, i.e. respondents could lie both before and after the election. It is very difficult to rule out this possibility *definitively*, but indirect evidence suggests that it is unlikely. In particular, the two post-election random-probability surveys—the BES and the BSA survey—produced estimates of the election result that were approximately correct (as discussed in Section 4.3), with both producing estimates of the Conservative vote share that were actually slightly above the result. We see no reason to assume that respondents should choose deliberately to misreport their past vote in some post-election surveys but not in others. In summary then, we rule out violations of assumption 3 such as late swing and deliberate misreporting, as having made any notable contribution to the polling miss.

4.2. Turnout weighting

The pollsters used a range of methods for constructing turnout weights p_{T_i}. Most relied on responses to a self-reported likelihood-to-vote (LTV) question such as 'how likely is it that you will vote in the general election on 7th May?', to which responses were recorded on scales of between four and 11 points. Some pollsters used the question as a binary filter (so that those below a threshold on the LTV question received a turnout weight of 0 and those above a weight of 1) and others in a smoother manner (e.g. by dividing a 0–10 LTV response by 10). Some turnout weights were based only on an LTV question, whereas others used additional information such as age or past voting. The models that were used to generate the turnout weights were educated guesses, with the exception of TNS UK who used a model that was fitted to data from the 2010 BES (which includes both an LTV question and a measure of validated vote).

Recall that assumption 2 requires that probabilities of voting in the election are allocated to the respondents on the basis of an accurate model for turnout, conditional on self-assessed likelihood of voting, L, voting intention V and auxiliary variables X. Specifically, the weights should accurately describe these probabilities in the population of voters. This presents a problem for assessing the adequacy of the turnout model, because this should ideally be done by using a high quality pre-election probability sample in which LTV, intended vote and turnout after the election are observed. Unfortunately, no such study was undertaken in 2015. What

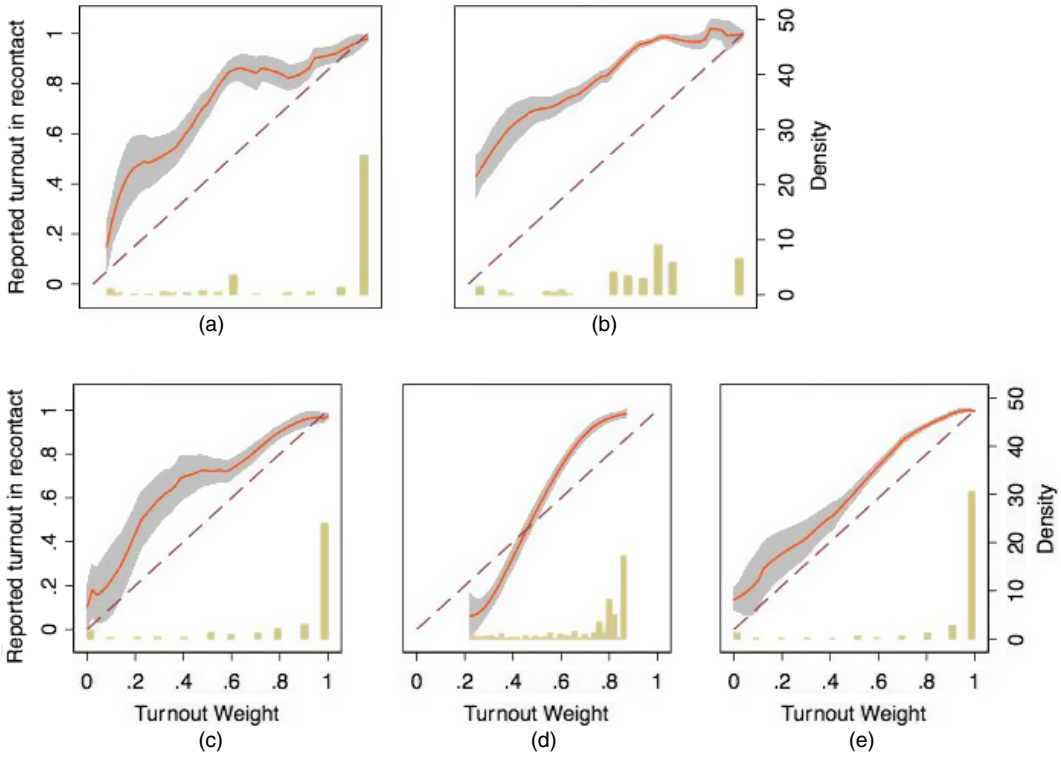

Fig. 1. Probability of turnout, estimated from five recontact surveys, as a smoothed function of turnout weights for the same respondents obtained from pre-election polls (———), with 95% confidence bands (▒) (the bar charts at the bottom of each plot show the relative frequencies of the values of the turnout weights): (a) ICM; (b) Populus; (c) Survation; (d) TNS UK; (e) YouGov

can be examined, though, is how well p_{T_i} approximated $p(T=1|V,L,X,S=1)$, i.e. the turnout probabilities in the poll samples. This is not conclusive evidence to make this assessment, because it requires the additional assumption that the model for these probabilities should be approximately the same for the poll respondents and the target population. The validity of this assumption cannot be directly assessed.

Fig. 1 provides information about the accuracy of the turnout weights as estimates of turnout probabilities for respondents in the five recontact surveys. The full curves show the probability of turnout as a smoothed function of the turnout weights, so the accuracy of the weights as probabilities of voting can be judged by the proximity of the full curves to the broken lines (on which the reported turnout rate is equal to the actual turnout weight). For all except one of the pollsters it is clear that actual turnout was higher, sometimes substantially higher, than the turnout weights implied, except where the weight was close to 1 (TNS UK was a partial exception here, producing turnout weights which were overestimates when the weight was less than 0.5). Some, though not all, of this inaccuracy in the turnout weights may be accounted for by overreporting of turnout in the recontact surveys.

The bar chart at the bottom of each plot in Fig. 1 shows the relative frequency of respondents with different values of the turnout weights. It is clear that a large majority of the respondents received a weight of 1. Almost all such respondents also reported after the election that they had voted. Thus, although the calibration of turnout weights was poor across the full range of

probabilities, for most of the poll respondents the weights were quite accurate because the vast majority reported that they would vote and they were allocated a turnout probability of 1.

The accuracy of the turnout weights are of little substantive interest in themselves. They matter only in so far as they affect estimated vote shares. Whether this was so for the 2015 polls may be assessed by calculating vote intention estimates under different specifications for the turnout weights. First, we can use the recontact polls to examine whether the estimated shares would have been different if turnout weights had not been needed at all, i.e. if the pollsters had known who would and would not turn out to vote. This is done by calculating estimates by using pre-election vote intention only for those respondents who are known (by self-report in the recontact surveys) to have voted in the election; these respondents can be assigned a turnout probability of 1. Estimates for the difference in the Conservative–Labour vote share by using this approach are shown in the first row of Table 3. They are between −2.1 and 0.5 percentage points, compared with from −2.7 to 0.7 points for the final polls (the latter for all nine BPC members). There is, thus, no evidence that the poll estimates would have been more accurate, even if the pollsters had known before the election which respondents would and would not turn out to vote.

We can also examine the sensitivity of vote intention estimates by calculating the party shares with various specifications for the turnout weights, keeping all other elements of the weighting unchanged. An example is presented in Fig. 2, which shows estimates of the Conservative lead for the final polls, with four different turnout weights (from left to right):

(a) using only those respondents who said that they were certain to vote, i.e. who gave the highest response to the LTV question;
(b) the turnout weights that were used for the published estimates;
(c) transformed weights ($p + p^*(1 - p)$ where p is the original turnout weight) which would have been closer to the true turnout probabilities in Fig. 1;
(d) giving every respondent a turnout probability of 1.

These quite different specifications do not change the estimates in any substantial way. We have also used a range of turnout weights from a model-based approach applied to the 2010 and 2015 BESs (of the kind used by TNS UK). These alternative probabilities also have no notable effect on the vote share estimates.

It is worth noting that none of the pollsters included vote intention in their models for turnout probability, so they implicitly assumed that the probability does not depend on the party that the respondent intends to vote for, once LTV and other variables have been controlled. If this assumption fails, supporters of one party would be more likely than another to vote, given their reported pre-election LTV. We refer to this possibility, which has the potential to cause biases in poll estimates of vote shares, as 'differential turnout misreporting'. We can assess whether this occurred in 2015 by including vote intention as a predictor in a model predicting turnout probability. Using this specification the party variable is statistically significant for only one pollster, and here the effect is in the opposite direction to what would be required to explain the polling miss; those who said that they intended to vote Labour were *more* likely to vote, given their answer to the LTV question. The recontact surveys therefore show no evidence of differential turnout misreporting.

In summary, there were notable inaccuracies in the turnout weights as estimates of actual turnout probabilities for the respondents to the 2015 election polls. However, this made little difference to the final vote shares; estimates of the Conservative lead would not have been more accurate, even if the turnout weights had been based on self-reported vote after the election, or if they had been assigned in a very different set of ways. Nor is there evidence that respondents who reported intending to vote Labour may have overestimated their future likelihood of turnout

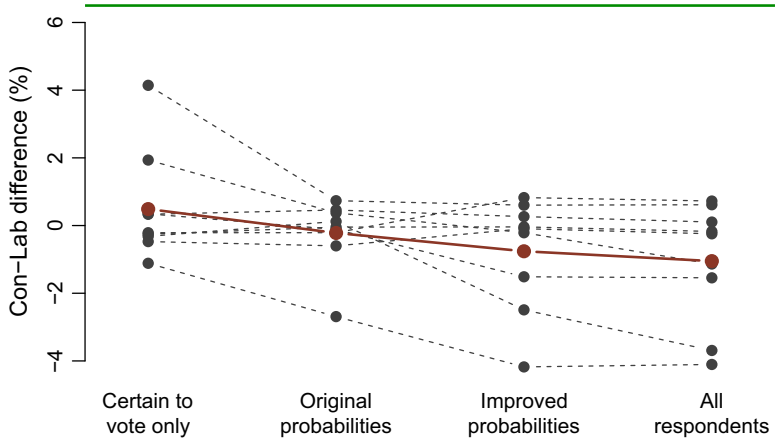

Fig. 2. Conservative lead over Labour (per cent for Great Britain), estimated from the final polls by the nine BPC members with various specifications of the turnout weights—from left to right, these specifications are as follows: using only the respondents who said that they were certain to vote, the original (pollster-specific) turnout weights p, weights of $p + p^*(1 - p)$ and using all respondents with weight 1, irrespective of stated likelihood to vote (——, true election result (6.5%); ——, unweighted average of the figures for the nine polls)

more than Conservative intenders. We conclude, on this basis, that violations of assumption 2 on turnout weighting were not responsible for the 2015 polling miss.

4.3. Representative sampling

We have concluded that violations of assumptions 2 and 3 relating to late swing and turnout weighting made little or no contribution to the 2015 polling error. By a process of elimination, then, we are led to conclude that violation of assumption 1—*representative sampling*—must have been the primary locus of the 2015 polling miss: the polls systematically overrepresented Labour voters and underrepresented Conservative voters in their weighted estimates.

In this section, we examine what direct evidence there is to support the judgement that the polling miss was due to unrepresentative samples. We first consider a comparison with two surveys that were undertaken shortly after the election and which used probability sampling designs: the BES and the BSA survey. We then examine estimates of vote shares by subgroups defined by the weighting variables, and then biases in estimates of other variables which are likely to be related to voting. Although none of these lines of evidence can be considered conclusive in themselves, collectively they provide consistent evidence to support the conclusion that the poll samples were systematically biased in their composition relative to the target population.

The BES and BSA employ what can be considered 'gold standard' procedures at all stages of their design but are most notably different from the polls in that they employ probability sampling rather than quota sampling. It is important to be clear that probability sampling does not on its own guarantee accuracy of survey estimates; these types of survey are themselves subject to various errors of observation and non-observation (Groves, 1989). In particular, when a substantial proportion of the eligible sample fails to complete the survey, either through refusal to participate or failure to be contacted, there is a risk that estimates will be biased because of differential non-response (although recent research has shown that the correlation between response rate and non-response bias is considerably weaker than has historically been assumed; see Groves and Peytcheva (2008) and Sturgis *et al.* (2016)). As we shall see, however, in 2015

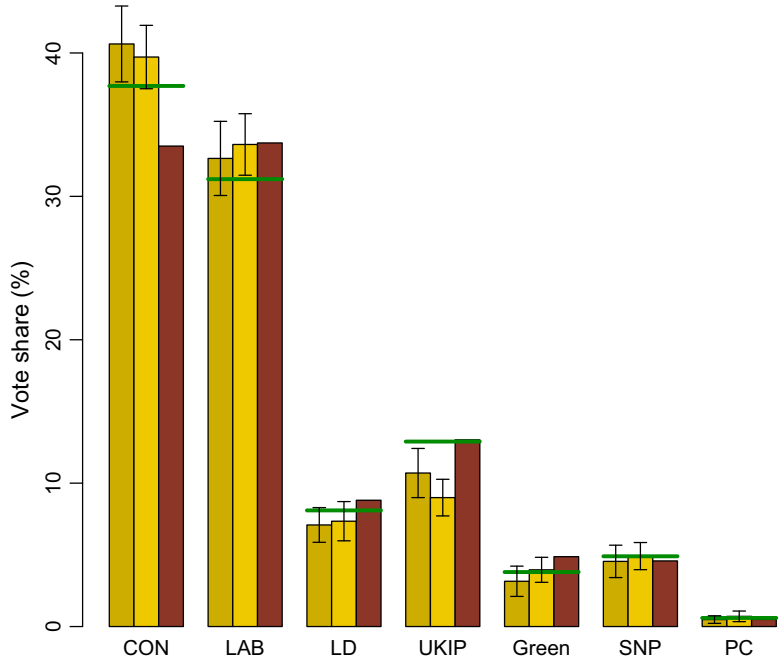

Fig. 3. Estimates of voting intention for the various parties (as percentages of vote in Great Britain): BES (■), BSA survey (■) and the average of the final polls by the nine members of the BPC (■) (——, election results; ⊥, 95% confidence intervals for BES and BSA survey estimates)

the BES and BSA survey were far more accurate than the pre-election polls in their estimates of the vote distribution and, given the transparency and robustness of their underlying sampling procedures, it is reasonable to use them as a lens through which to assess the quality of the poll samples which were obtained by using quite different approaches.

The reported vote distributions for the BES and BSA survey are shown in Fig. 3, alongside the average vote intention estimates for the final polls and the election result. It is immediately apparent that the BES and BSA survey produced more accurate estimates of the Conservative lead over Labour than the polls, with the BES showing a 7-point lead and the BSA survey a 6-point lead for the Conservatives. Neither of these surveys was itself completely accurate, with both significantly underestimating the UK Independence Party share, the BES overestimating the Conservative share and the BSA survey overestimating the Labour share.

This comparison is suggestive that the polls underestimated the Conservative lead as a result of their sampling procedures. However, it is inconclusive on this point because the BES and BSA survey differ from the polls in other respects, beyond their sample designs. Most importantly, both were undertaken *after* the election had taken place. This means that there was no uncertainty (at least by self-report) about whether the respondents had voted or not when they reported their vote choice, whereas the polls had to factor in whether a respondent would actually vote or not to their pre-election estimates. The reported votes of the BES or BSA survey respondents might also have been influenced by their knowledge of the election result, which could not have been so for the pre-election polls. Previous research has shown a tendency for respondents to recall disproportionately having voted for the winning party—so called 'bandwagoning' (Nadeau *et al.*, 1993) and such effects might plausibly have contributed to the difference in the lead estimates between the surveys and the polls in 2015 (we note that

this might also have affected the recontact polls that were discussed in Section 4.1, in which case bandwagoning would have exaggerated the before–after differences that are reported in Table 3).

Another potentially consequential difference is the mode of interview, with the BES and BSA survey using face-to-face interviews and the polls using either telephone interviews or on-line self-completion. Although face-to-face interviewing is generally acknowledged to be the gold standard for survey modes (Couper, 2011), this does not imply that it will produce more accurate self-reports of vote choice than the other modes. Indeed, the survey methodological literature suggests that the face-to-face interviewing is more prone to measurement error due to socially desirable responding than telephone and self-completion modes (Tourangeau *et al.*, 2000). Nonetheless, these factors all render the headline comparison between the polls and BES and BSA survey ambiguous with regard to the underlying cause of the difference.

Fortunately, we can effectively rule out the two most important of these design differences by considering the reported vote distributions for the polls that undertook recontact surveys. Because the recontact surveys were carried out after the election, we can exclude timing relative to the election as a potential confounder. Table 3 shows that there is only a very modest improvement in poll estimates (weighted for attrition) of the Conservative lead when the polls are undertaken after the election and respondents are reporting their actual, rather than their intended, vote. These comparisons, then, support the conclusion that the differences between the BES and BSA survey and the polls were due to differences in their sampling procedures, rather than to whether they were undertaken before or after the election. A *caveat* to this conclusion is that the fieldwork periods were much shorter for the recontact polls than for the BES and BSA survey, so bandwagoning may have been more prevalent in the latter than the former case. However, Mellon and Prosser (2017) demonstrate that this possibility has little empirical support, for the BES at least.

Recall that assumption 1 for representative sampling requires that, for any given value of the weighting variables X, observations (V_i, L_i) in a poll are a random sample from $p(V, L|X)$ in the population. It is informative, therefore, to assess the extent to which the polls were in error not only in the aggregate but also across the weighting cells that were used by the pollsters. Fig. 4 presents estimates of the Conservative–Labour difference by exemplar weighting variables, compared with the actual election result (for region) or with estimates from the BES and BSA survey (combined where both are available, because of small sample sizes by weighting cells for each survey on its own). It can be seen that there is no apparent difference in the polling error between men and women. When considered by age band, however, the polls substantially underestimate the Conservative lead among those aged 45–64 years and, to a lesser extent, those aged 55–64 years. Here, of course, we must assume that the BES or BSA survey distribution is approximately correct within age bands, although this does not seem unreasonable given that both surveys calculated the population estimate of the Conservative lead approximately correctly.

Considered by self-reported vote in the 2010 general election, the pattern in Fig. 4 suggests that the polls were most inaccurate for those who voted for the two main parties in 2010. Finally, at the government office region level the results suggest that the polls particularly underestimated the Conservative lead in regions where the Conservative vote share was higher than the national average; the East, East Midlands, South West and South East. In sum, these analyses clearly demonstrate that the key assumption of representativeness of vote intention within weighting cells was strongly and consistently violated in the 2015 polls. The pattern that we observe in these charts also suggests a systematic tendency for the polls to underestimate Conservative support in subgroups where the Conservative lead over Labour is highest, e.g. older people, southern counties and people who voted Conservative in 2010. We cannot pursue this further empirically

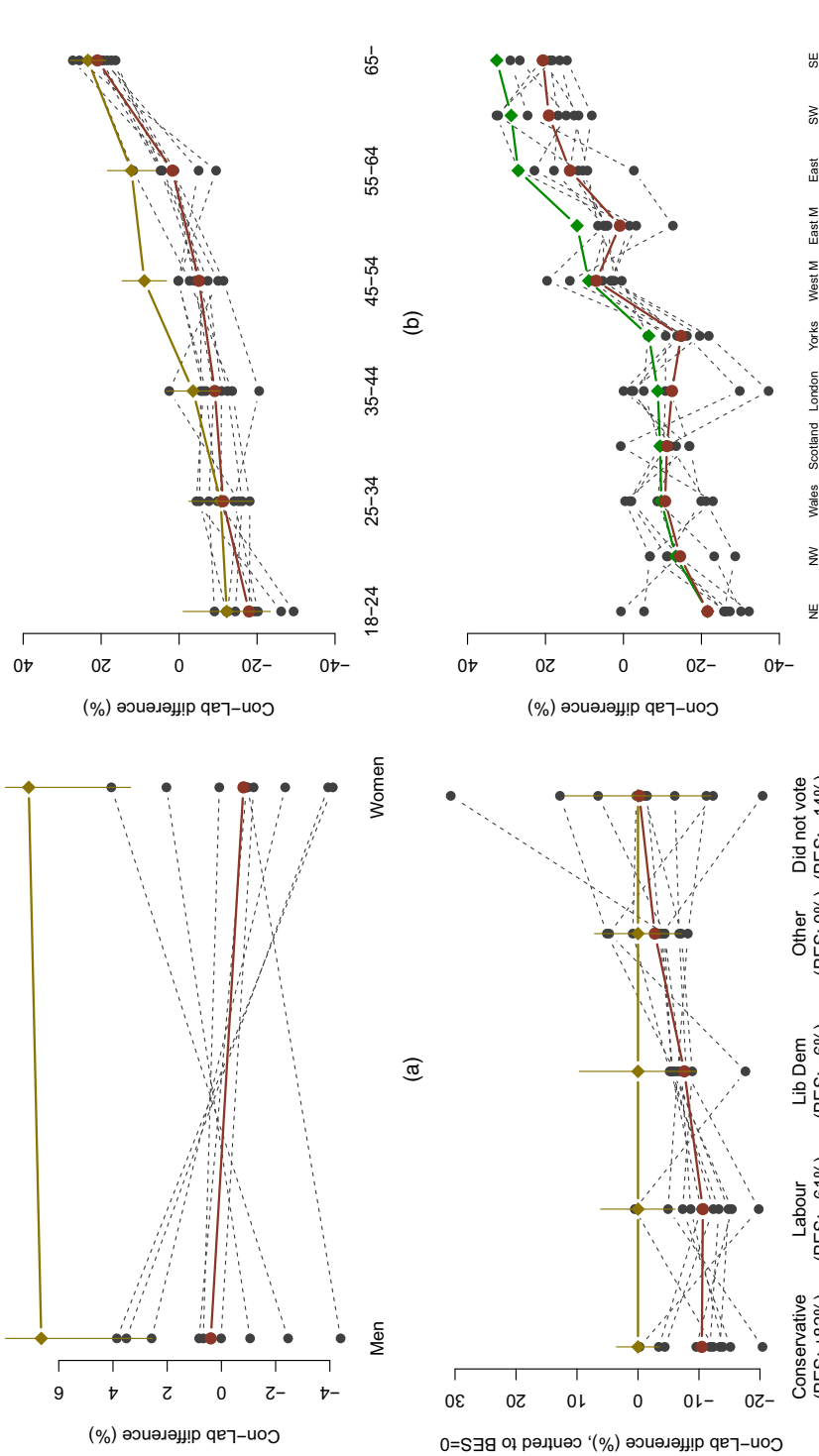

Fig. 4. Conservative lead over Labour (per cent for Great Britain), estimated from nine final polls, separately by categories of some weighting variables (●, average of the polls; ◆, estimate from the combined BES–BSA survey sample, with 95% confidence intervals (2010 vote from the BES only); ◆, 2015 election result) (in the plot for the 2010 vote, the vertical axis is centred so that 0 represents the difference in the BES (which is shown on the horizontal axis; for example, among those who reported voting Conservative in 2010, the Conservative–Labour difference in the BES was 82%)): (a) gender, (b) age group, (c) region and (d) party vote in the 2010 general election

given the data that were available to us, but it seems likely that a key reason that the polls underestimated the Conservative lead over Labour is that their sampling procedures systematically under-represented Conservative voters within these kinds of Conservative supporting demographic groups.

A third type of comparison is informative about the representativeness of the poll samples; how accurate the poll estimates are for other variables that were measured in the polls and which are themselves related to vote choice. Consider, for example, sector of employment; it is known that, broadly, public sector workers are more likely to vote Labour and private sector workers are more likely to vote Conservative (Dunleavy, 1980). If polls that do not weight to population totals for employment sector were found to have overestimated the proportion of voters who work in the public sector, then this would not only constitute evidence that the poll samples were unrepresentative with regard to employment sector; it would also suggest a potential cause of the bias in the vote choice estimate. That is to say, by over-representing public sector workers in their samples, the polls would have overestimated support for Labour and underestimated support for the Conservatives. This approach is appealing because it indicates ways in which poll samples might be improved in the future, either through changes to sample recruitment procedures or through improvements to quota and weighting targets (see for example Mellon and Prosser (2017) and Rivers and Wells (2015)).

Unfortunately, the extent to which we can implement this strategy is constrained by the paucity of candidate variables in the poll samples for which gold standard estimates are also available. Variables which meet these twin criteria are, almost by definition, scarce. If they had been collected, the pollsters would probably already be using them in their sampling and weighting procedures. Nonetheless, some variables are available which enable us to consider the polls from this perspective, albeit in a more limited manner than we would ideally like.

The first example relates to the continuous age distribution within banded age ranges. All pollsters weight their raw sample data to match the distribution of age by banded ranges in the population census. Three of the BPC members also recorded continuous age, making it possible for us to assess the age distribution within age bands and to compare this with the distribution from the census and the BES or BSA survey (the three BPC member polls were all conducted on line, so we cannot conclude that the same effect is apparent in phone samples). Fig. 5 displays

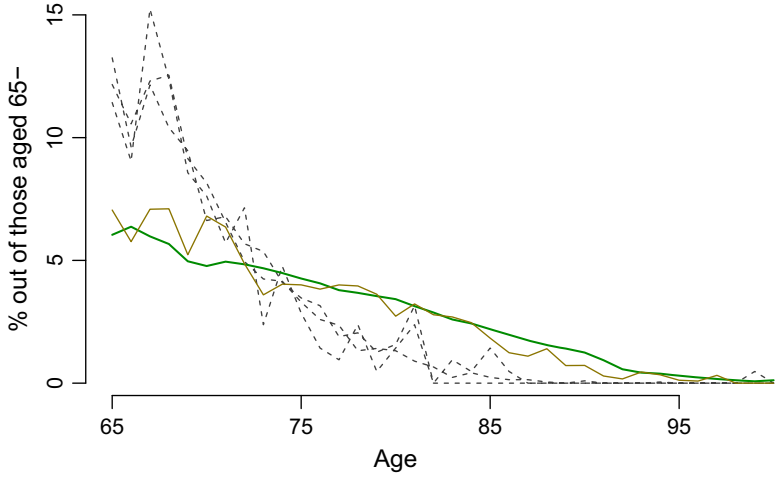

Fig. 5. Age distribution of the respondents aged 65 years or over in three final polls, compared with the distribution in the BES and the BSA survey (——, combined sample) and in the 2011 census (——)

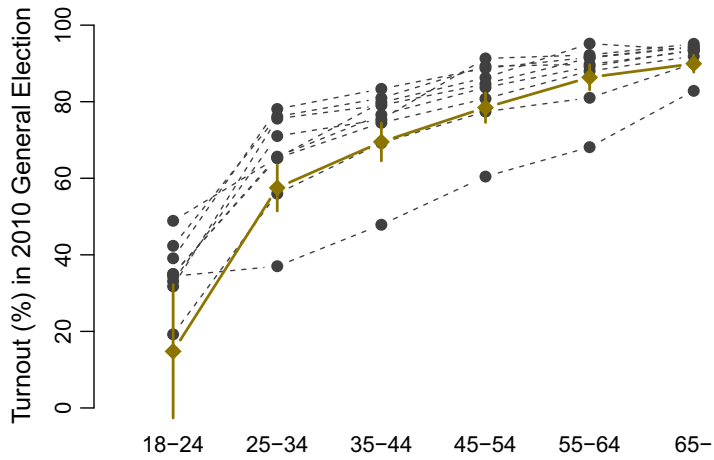

Fig. 6. Proportion of individuals who report that they voted in the 2010 general election, by age band (in 2015), estimated from nine final polls, compared with the estimates from the 2015 BES (———), with 95% confidence intervals

this comparison for the oldest age band: those aged 65 years and older. It shows that the polls substantially over-represent people under the age of 70 years and under-represent those aged 75 years and older within this age band, whereas the BES and BSA survey do not. Indeed, the three polls that are included here contain almost no respondents aged 90 years or above. It is easy to see how this kind of imbalance could arise in the quota sampling that is used for the polls, where the oldest age group for the quota targets was typically 65 years and older. If individuals towards the younger end of this age band were easier to include in the poll, the 'oldest old' would end up being under-represented in the sample.

This is itself direct evidence that poll samples can produce quite biased estimates of population characteristics. However, it also indicates the kinds of selection mechanism which might, in part, have led to the 2015 polling miss. If the Conservative lead over Labour was bigger among voters aged over 74 years than those aged between 65 and 74 years, then under-representing the older age group would have biased the estimate of the Conservative lead towards 0. In fact, the 2015 BES shows that the Conservatives held a 21-point lead over Labour among those aged over 74 years and a 22-point lead among those aged 64–74 years. So the under-representation of voters aged 75 years and over in the poll samples seems unlikely to have made a notable contribution to the 2015 polling miss.

A second example of biased estimates in the poll samples relates to reported turnout in the 2010 general election. Fig. 6 plots self-reported 2010 turnout by age band for the 2015 polls and for the BES. With one exception, the polls consistently overestimate turnout in the 2010 election (even compared with the BES, where turnout may also be over-reported), with a particularly large bias among those aged 18–24 years. Given that only around a third of this cohort would even have been eligible to vote in 2010, these are very substantial overestimates of the true proportion. A similar pattern has been observed in pre-election polls for other indicators of political engagement (Rivers and Wells, 2015; Mellon and Prosser, 2017).

In Section 2 we identified assumption 1 of representative sampling to be particularly strong. In this section we have assessed the empirical evidence that this assumption was violated in the 2015 polls. We have shown that estimates from surveys using random-probability sampling produced accurate estimates of the Conservative lead over Labour (and that this difference cannot be attributed to their having been undertaken after the election), that the polls exhibited

substantial biases within weighting cells and that biases were evident on other variables in addition to vote intention. Individually and collectively these findings support the conclusion that unrepresentativeness of the poll samples on vote intention was the key contributory factor in the 2015 polling miss.

5. Discussion

In the months and weeks leading up to the 2015 general election the polls told a consistent story: the Conservatives and Labour were in a dead heat in the popular vote. This led media commentators, party strategists and the public to focus attention on the likely composition of a coalition, rather than on a single-party government led by the Conservatives who, of course, ultimately won the election with a 6.5% lead over Labour and an absolute majority in the House of Commons. The expectation of a hung Parliament in the final days and weeks of the campaign was so strong and widely held that the sense of shock and disbelief was palpable when the result of the exit poll was announced at 10 p.m. on May 7th.

Having considered a range of plausible contributory factors and sources of data, our analyses lead us to conclude that the primary cause of the polling miss was that the samples were unrepresentative of the population of voters. In short, the methods that were used to collect samples of voters systematically over-represented Labour supporters and under-represented Conservative supporters. The statistical adjustment procedures that were applied to the raw data did not mitigate this basic problem to any notable degree.

We came to this conclusion partly by elimination of other putative causes of the error. The discrepancy between the point estimates of the Conservative lead in the polls and the election result cannot be attributed to sampling error. Using a new procedure for calculating the precision of vote share estimates from quota samples, we have shown that none of the BPC pollsters' estimates contained the true lead of the Conservatives over Labour within the 95% confidence interval. We recommend that pollsters move to adoption of this, or a similar, approach to estimating the sampling variance of party vote shares. We could also replicate all published estimates for the first, the penultimate and the final published polls by using the raw microdata provided by the BPC pollsters, enabling us to rule out the possibility that some of the errors were due to flawed analysis or use of inaccurate weighting targets.

We found some evidence that there may have been a modest shift in vote share towards the Conservatives between the final polls and election day, although this can have contributed at most around one and a half percentage points to the mean error on the Conservative lead. The widely held view that the polling miss was due to deliberate misreporting—'shy Tories' telling pollsters that they intended to vote for other parties—is very difficult to reconcile with the results of the recontact surveys that were carried out by the pollsters and with the BES and BSA survey undertaken after the election by using random-probability sample designs. Ruling out this kind of shift also enables us to discount measurement error arising from question wording and order as a possible cause, because this is a special case of the same overarching phenomenon.

Differential turnout was also pointed to after the election as a likely cause of the errors; so-called 'lazy Labour' supporters telling pollsters that they would vote Labour but ultimately not turning out to vote. Data from various sources show no support for differential turnout misreporting, or errors in predicted probabilities of turnout in general, making anything except a very small contribution to the polling errors. This means that we can also reject the possibility that unregistered voters made any contribution to the polling errors because this would manifest as an error of turnout weighting.

If the potential causes that were considered above are found to have made, at best, marginal contributions to the polling error, we are left to conclude that unrepresentativeness in the samples must have been the primary cause of the polling miss in 2015. On its own, a strategy which reaches a conclusion through elimination of alternative explanations is not very satisfactory, particularly when the evidence on which the preliminary eliminations are based is imperfect, as here. If we had been drawn, by a process of elimination, to conclude that the polling miss was due to a *prima facie* implausible explanatory factor—such as overseas voters—then we would question the validity of the process that led us to this inference. But this is not so here; we identified sampling and weighting procedures as representing inherent weaknesses in our description of the assumptions underlying the methodology of polling.

We have also provided empirical evidence in support of the conclusion that the sampling procedures that were employed by the pollsters produced biased estimates of vote intentions. Random-probability samples undertaken shortly after the election produced accurate estimates of the Conservative lead over Labour, suggesting that the less robust sampling procedures that were used by the polls were responsible for the underestimation of this key parameter. The difference in the estimate of the Conservative lead between the probability samples and the polls is still evident by using the recontact surveys that were undertaken by a subset of pollsters, indicating that the sampling procedures rather than the timing of the fieldwork were the cause of the difference in the estimates of the lead. Additionally, we showed that the polls strongly violated the core assumption that is required for representative sampling; that the estimates of vote intention should be accurate conditional on the variables that are used for calibration weighting. It was particularly suggestive that the polls underestimated the Conservative lead most in areas and subgroups where the true Conservative lead was largest. Finally, we presented specific examples of two other variables in the poll samples, age and turnout in the 2010 election, on which biases were also evident. Taken together, these findings lead us to conclude that violation of the representative sampling assumption was the primary cause of the 2015 polling miss.

What can be done to improve the representativeness of poll samples in the future? The answer to this question depends on whether the pollsters continue to employ quota methods, or switch to random-probability sampling. Because of the high cost of probability sampling, we expect the vast majority of opinion polls to continue to use non-random sampling methods for the foreseeable future. The aim then must be to do them as well as possible: a point that is not limited to election polls but refers to the increasing volume of research being undertaken using on-line non-probability samples (for an introduction to the growth of non-probability sampling, see the American Association for Public Opinion Research report by Baker *et al.* (2013), and its references and associated discussion).

Continuing with non-random sampling in the election polls means that there are only two broad strategies that can be pursued to improve sample representativeness. Pollsters can take measures to increase the representativeness of respondents who were recruited to existing quota and weighting cells, or they can incorporate new variables in their weighting schemes which are related to both the probability of selecting into poll samples and vote intention. These are not mutually exclusive strategies.

How this is done will depend, to an extent, on the mode of interview of the poll. For phone polls this is likely to involve (but will not be limited to) using longer fieldwork periods, more call-backs to initially non-responding numbers (both non-contacts and refusals) and ensuring a more representative mix of landline and mobile phone numbers. We recognize that, taken to their logical extreme, these procedures would be practically equivalent to implementing a random-probability design and would therefore be expensive and time consuming. Although, as we shall note shortly, we would very much welcome the implementation of truly random-sample designs,

we acknowledge that the cost restrictions of true random methods make them impractical for the vast majority of pre-election phone polls. The extended fieldwork periods that are required for high quality random samples also means that they have obvious weaknesses for campaigns characterized by volatile voter preferences. Nevertheless, it would seem that there are gains to be made in quality without making the resultant design unaffordably expensive and lengthy. It may be that implementing procedures of this nature results in fewer polls being carried out than in the last Parliament, as the cost of undertaking each would no doubt increase. This would, in our view, be no bad thing, so long as the cost savings that accrue from doing fewer polls are invested in quality improvements.

For on-line polls the procedures that are required to yield more representative samples within weighting cells are also likely to involve longer field periods and more reminders, as well as differential incentives for under-represented groups, and changes to the framing of survey requests. We encourage on-line pollsters to experiment with these and other methods to increase the diversity of their respondent pools.

The second strategy that pollsters can pursue to improve sample representativeness is to modify the variables that are used for the weighting of the poll data. Here, there is not such a clear trade-off between expense and quality as there is with obtaining more representative samples. If variables that are correlated with self-selecting into opinion polls and vote intention were readily available, pollsters would already be using them. We also recommend caution in the use of variables for calibration weighting which do not have well-defined and correctly known population totals.

Despite its limitations, polling remains the primary means of estimating likely vote shares in elections and this, we contend, is likely to remain so for the foreseeable future. Although polls rarely produce exactly correct vote share estimates and sometimes produce substantial errors, the historical record shows that the final polls are usually within a few percentage points of the actual party shares in elections (Sturgis *et al.*, 2016). Yet, it must be better acknowledged that accurately predicting vote shares in an election is a very challenging task (Hillygus, 2011). A representative sample of the general population must be obtained and accurate reports of party choice elicited from respondents. An approximately accurate method of determining how likely respondents are to cast a vote must be implemented and the sample of voters must not change their minds between taking part in the poll and casting their ballots. What is more, the entire procedure must usually be carried out and reported on within a very short space of time and at very low cost. Given these many potential pitfalls, it should come as no surprise that the historical record shows polling errors of the approximate magnitude of 2015 occur at not infrequent intervals.

Acknowledgements

We are grateful to John Mellon and Chris Prosser for undertaking analyses using BES data and to Jack Blumenau and Rosie Shorrocks who provided research support. Any errors or omissions are our own.

We also acknowledge the support of the Economic and Social Research Council (National Centre for Research Methods grant RES-576-47-5001).

References

American Association for Public Opinion Research (2009) An evaluation of the methodology of the 2008 pre-election primary polls. *Report*. American Association for Public Opinion Research, Chicago. (Available from

https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Rept_FINAL-Rev-4-13-09.pdf.)
American Association for Public Opinion Research (2017) An evaluation of 2016 election polls in the United States. *Report*. American Association for Public Opinion Research, Chicago. (Available from http://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx.)
Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013) Summary report of the AAPOR task force on non-probability sampling. *J. Surv. Statist. Methodol.*, **1**, 90–143.
Butler, D. and Pinto-Duschinsky, M. (1971) *The British General Election of 1970*. London: Macmillan.
Clery, E., Curtice, J. and Harding, R. (2016) British social attitudes: the 34th report. *Report*. National Centre for Social Research, London. (Available from www.bsa.natcen.ac.uk.)
Converse, J. (1987) *Survey Research in the United States: Roots and Emergence 1890–1960*. Berkeley and Los Angeles: University of California Press.
Couper, M. P. (2011) The future of modes of data collection. *Publ. Opin. Q.*, **75**, 889–908.
Crossley, A. (1937) Straw polls in 1936. *Publ. Opin. Q.*, **1**, Jan., 24–35.
Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
Deming, W. E. and Stephan, F. F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, **11**, 427–444.
Deville, J.-C. (1991) A theory of quota surveys. *Surv. Methodol.*, **17**, 163–181.
Deville, J.-C. and Särndal, C.-E. (1992) Calibration estimators in survey sampling. *J. Am. Statist. Ass.*, **87**, 376–382.
Dunleavy, P. (1980) The political implications of sectoral cleavages and the growth of state employment: Part 1, the analysis of production cleavages. *Polit. Stud.*, **28**, 364–383.
Fieldhouse, E., Green, J., Evans, G., Schmitt, H., van der Eijk, C., Mellon, J. and Prosser, C. (2015) British Election Study, 2015: face-to-face survey (computer file). University of Manchester, Manchester. (Available from http://www.britishelectionstudy.com/bes-resources/f2f-v1-0-release-note/.)
Groves, R. M. (1989) *Survey Errors and Survey Costs*. New York: Wiley.
Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009) *Survey Methodology*, 2nd edn. Hoboken: Wiley.
Groves, R. M. and Peytcheva, E. (2008) The impact of nonresponse rates on nonresponse bias. *Publ. Opin. Q.*, **72**, 167–189.
Hawkins, O., Keen, R. and Nakatudde, N. (2015) General Election 2015. *Briefing Paper CBP7186*. House of Commons Library, London. (Available from http://researchbriefings.files.parliament.uk/documents/CBP-7186/CBP-7186.pdf.)
Hillygus, D. S. (2011) The evolution of election polling in the United States. *Publ. Opin. Q.*, **75**, 962–981.
Keeter, S., Igielnik, R. and Weisel, R. (2016) Can likely voter models be improved?: Evidence from the 2014 U.S. House Elections. Pew Research Center, Washington DC. (Available from http://www.pewresearch.org/files/2016/01/PM_2016-01-07_likely-voters_FINAL.pdf.)
Kish, L. (1965) *Survey Sampling*. New York: Wiley.
Market Research Society (1994) *The Opinion Polls and the 1992 General Election*. London: Market Research Society.
Mellon, J. and Prosser, C. (2017) Missing non-voters and misweighted samples: explaining the 2015 Great British polling miss. *Publ. Opin. Q.*, **81**, 661–687.
Mosteller, F., Hyman, H., McCarthy, P., Marks, E. and Truman, D. (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.
de Munnik, D., Illing, M. and Dupuis, D. (2013) Assessing the accuracy of non-random business conditions surveys: a novel approach. *J. R. Statist. Soc.* A, **176**, 371–388.
Nadeau, R., Cloutier, E. and Guay, J.-H. (1993) New evidence about the existence of a bandwagon effect in the opinion formation process. *Int. Polit. Sci. Rev.*, **14**, 203–213.
Rivers, D. (2013) Comment on 'Summary report of the AAPOR task force on non-probability sampling'. *Surv. Statist. Methodol.*, **1**, 111–117.
Rivers, D. and Wells, A. (2015) Polling error in the 2015 UK General Election: an analysis of YouGov's pre and post-election polls. YouGov UK, London. (Available from https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/x4ae830iac/YouGov%20%E2%80%93%20GE2015%20Post%20Mortem.pdf.)
Rust, K. F. and Rao, J. N. K. (1996) Variance estimation for complex surveys using replication techniques. *Statist. Meth. Med. Res.*, **5**, 283–310.
Smith, T. M. F. (1983) On the validity of inferences from non-random samples. *J. R. Statist. Soc.* A, **146**, 394–403.
Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. and Smith, P. (2016) Report of the inquiry into the 2015 British General Election opinion polls. *Report*. Market Research Society and

British Polling Council, London. (Available from http://eprints.soton.ac.uk/390588/1/Report_final_revised.pdf.)
Tourangeau, R., Rips, L. and Rasinski, K. (2000) *The Psychology of Survey Response*. New York: Cambridge University Press.
Traugott, M., Bolger, G., Davis, D. W., Franklin, C., Groves, R. M., Lavrakas, P. J., Mellman, M. S., Meyer, P., Olson, K., Selzer, J. A. and Wlezien, C. (2009) An evaluation of the methodology of the 2008 pre-election primary polls. American Association for Public Opinion Research Ad Hoc Committee on the 2008 Presidential Polling. American Association for Public Opinion Research, Chicago.
Wolter, K. M. (2007) *Introduction to Variance Estimation*, 2nd edn. New York: Springer.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'An assessment of the causes of the errors in the 2015 UK General Election opinion polls: Supplementary materials'.