Effect of Weighting Poll Data to 2014 Indyref Vote on Reported Pro-independence Voting Intention

Mark McGeoghegan

Wednesday 21 December 2022

TL;DR

- The methodologies of research agencies conducting Scottish independence polling have come under greater-than-normal scrutiny in the past few weeks, as a result of a glut of polls showing support for Scottish independence at a higher level than support for the Union. These polls immediately followed a UK Supreme Court ruling that the Scottish Parliament cannot unilaterally hold a referendum on Scottish independence, and have driven a narrative that this ruling (among other factors) has led to a rise in support for independence.
- In particular, campaigners have made the claim that companies that do not weight their data to reflect the outcome of the 2014 Scottish independence referendum report higher pro-independence voting intention than those that do. Some pro-Union campaigners have characterised this as an unfair bias that produces inaccurate results.
- We aim here to determine whether or not there is a **statistically and substantively significant difference** in reported pro-independence voting intention between polls that do and do not weight by 2014 vote, by developing a robust multivariate regression model.
- We find that there is a **statistically significant difference in reported pro-independence voting intention** between polls that do and do not weight by 2014 vote, but that **the effect size of this difference is relatively small - 1.86pts**.
- We further find that whether or not a poll weights to the Scottish vote in the most recent UK Parliament election has a statistically significant relationship with reported pro-independence voting intention, of 1.35pts, as does excluding 16- and 17-year-olds - 1.46pts.
- Whether or not research agency *should* weight by 2014 vote is beyond the scope of this write-up

Background

Secessionism in Scotland is a highly sensitive topic. The Scottish National Party's dominance of Scottish politics since 2011 has produced three pro-independence Scottish Parliaments (since 2016, in conjunction with the Scottish Greens), and three pro-independence majorities of Scottish seats in the UK Parliament. It also produced a referendum on Scottish independence in 2014, which was lost by the pro-independence camp by 55% of the vote to 45%.

Consistent pro-independence majorities of seats in a country that voted by majority to remain in the UK has kept secession high up the political agenda without actually resolving the issue. As time has passed and positions have entrenched on both sides, being 'Yes' (pro-independence) or 'No' (pro-Union) has become a political identity, with strong polarisation between the two poles of opinion.

Accordingly, coverage of secessionism and political campaigning for and against it is often received through a partisan prism - online advocates for both sides are particularly prone to react to coverage and commentary by amplifying content that they perceive to benefit their own camp, and by undermining content that is detrimental to their cause. This partisanship often takes the form of conspiracy theory, and even ad hominem attacks.

In this context, opinion polling on Scottish independence has become politicised in the same manner as any other content. Activists and campaigners on both sides are prone to accusations of bias or dodgy methodology towards research agencies the results of whose polls they do not like.

In recent weeks, this contestation over opinion polling has intensified. The UK Supreme Court ruled in November that the Scottish Parliament and Government are not, under the Scotland Act 1998, empowered to hold a referendum on Scottish independence without the UK Government consenting to one. In the aftermath, a series of six polls have found a rise in support for Scottish independence of around 4-5 percentage points. Some polls found larger rises and higher overall levels of support for independence than other polls.

Several of these polls have come under attack over their weighting schemes, in particular because three of the post-UKSC ruling polls do not weight data by vote in the 2014 independence referendum - which some campaigners claim 'inflates' support for independence.

Aims

The purpose of this model is to assess whether or not weighting a poll by 2014 vote has an impact on the reported level of support for independence in a given poll. It is *not* its purpose to argue that it is right or wrong to do so. There are arguments for (to more accurately reflect the stated preferences of an electorate) and against (false recall and demographic churn since 2014), which I will not litigate here.

Thus far, campaigners claiming that not weighting by 2014 vote has a significant impact on reported pro-independence voting intention have supported this claim with side-by-side comparisons of selected polls which do and do not weight by 2014 vote.

There are obvious flaws with this kind of analysis. Firstly, the selected polls are not necessarily representative of Scottish independence polling as a whole.

Secondly, no attempt is made to test whether such differences are statistically significant or as likely to be a result of statistical noise as they are to be the result of structural differences between polls.

And thirdly, such analysis omits other variables - other weighting factors, recruitment method, data collection method, sample structure, method of likelihood-to-vote calculation - which vary structurally between research agency.

The aim here is to begin to address these defects by:

1. Conducting an analysis of *all* Scottish independence voting intention polls since the 2014 Scottish independence referendum.
2. Carrying out relevant statistical testing to determine whether there is a statistically significant difference between polls based on 2014 vote weighting.
3. Developing a regression model that includes other variables one might expect to influence the outcome of a poll, within the bounds of the information available about these polls.

Scottish independence polling since 2014

Between October 2014 and December 2022, 204 polls were conducted which asked representative samples of Scottish voters whether Scotland should be an independent country, using the question wording from the 2014 referendum ballot and a Yes/No/'Don't know' answer scale, and which included some form of likely voter modelling.

Table 1: Descriptive Statistics (Continuous Variables)

	Minimum	Maximum	Mean	Median	Standard Deviation
Yes	0.34	0.55	0.44	0.44	0.03
No	0.37	0.56	0.47	0.47	0.03
Undecided	0.010	0.220	0.083	0.080	0.030

Source: Scottish Independence Voting Intention Polls (October 2014 - December 2022)

These polls form the sample for this analysis. *Table 1* summarises the support of and opposition to Scottish independence found across these polls.

Table 2 summarises the rest of the dataset used in this analysis. Almost 7-in-10 (68%) of polls conducted by 2014 have been by three research agencies - Panelbase, Survation, and YouGov. All three of these agencies weight their data by 2014 vote. The only other agencies to have conducted polling into the double-digits are Savanta and Ipsos. Most, but not all, Savanta polls are weighted using 2014 vote. No Ipsos poll is weighted to 2014 vote.

As a result of the preponderance of polls being conducted by a handful of agencies, which for the most part weight by 2014 vote, 4-in-5 (81%) polls are weighted in this way. Two-in-five (39%) polls not weighted by 2014 vote were conducted by a single agency, Ipsos.

The vast majority (89%) of polls were conducted via an online panel methodology. 19 - including Ipsos' 15 - were conducted by telephone. Just 3 were conducted face-to-face, all by TNS.

Table 2: Descriptive Statistics (Categorical Variables)

	Count (%)
Research agency that conducted fieldwork	
BMG Research	8 (3.9%)
Deltapoll	1 (0.5%)
FindOutNow	2 (1.0%)
Hanbury Strategy	2 (1.0%)
ICM	1 (0.5%)
Ipsos	15 (7.4%)
JL Partners	1 (0.5%)
Lord Ashcroft Polls	2 (1.0%)
Opinium	5 (2.5%)
Panelbase	55 (27%)
Redfield and Wilton	3 (1.5%)
Savanta	22 (11%)
Stack Data	2 (1.0%)
Survation	42 (21%)
TNS	3 (1.5%)
YouGov	40 (20%)
Method of data collection	
Telephone	19 (9.3%)
Online Panel	182 (89%)
Face-to-face	3 (1.5%)
Scottish election franchise (16+) or UK election franchise (18+)	
18+	44 (22%)
16+	160 (78%)

Table 2: Descriptive Statistics (Categorical Variables) *(continued)*

	Count (%)
Past Vote Weights	
Weighted to 2014 independence referendum vote	166 (81%)
Weighted to constituency vote in the most recent Scottish Parliament election	70 (34%)
Weighted to vote in the most recent UK Parliament election	131 (64%)

Source: Scottish Independence Voting Intention Polls (October 2014 - December 2022)

Comparisons of 2014 Vote Weighted and Non-Weighted Polls

Our question of interest is whether or not weighting poll data to reflect the outcome of the 2014 independence referendum has a statistically and substantially significant effect on reported pro-independence voting intention in that poll. We can formalise this in order to test it, in the form of the following hypothesis:

H1: Scottish independence voting intention polls with data weighted to reflect the outcome of the 2014 Scottish independence referendum will report *lower* pro-independence voting intention than polls that are not so weighted.

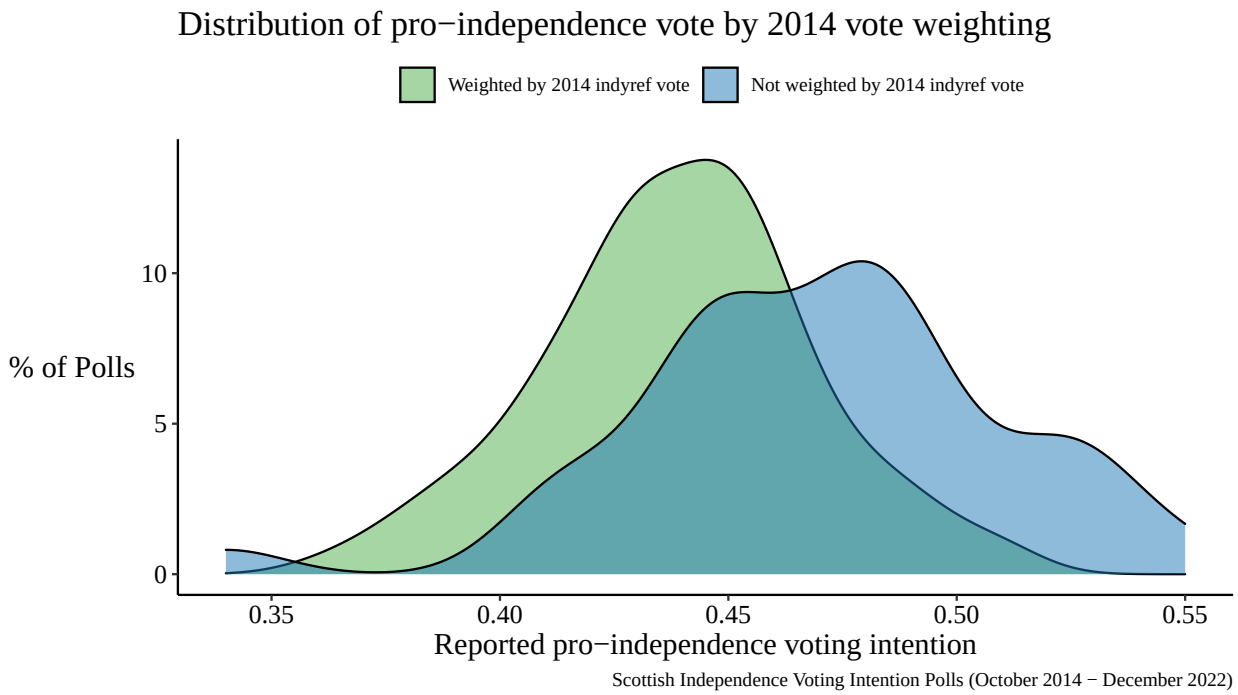
So, is there a difference? As we can see in *table 3*, polls that weight data by 2014 have a lower mean pro-independence voting intention (43.9%) compared to those that do not weight by 2014 vote (47%) - a difference of 3.1pts. A smaller gap exists for mean pro-Union voting intention (1.6pts) and undecideds (1.3pts). We see a similar pattern for median voting intention.

Table 3: Voting Intention Descriptive Stats by Weighting

	Not Weighted to 2014 Vote			**Weighted to 2014 Vote**		
	Mean	Median	Std Dev	Mean	Median	Std Dev
Yes	0.47	0.47	0.04	0.439	0.440	0.029
No	0.454	0.450	0.030	0.47	0.47	0.03
Undecided	0.073	0.060	0.029	0.086	0.080	0.030

Source: Scottish Independence Voting Intention Polls (October 2014 - December 2022)

However, we can also see from *table 3* that the standard deviations for mean voting intention in both groups of polls are as large, if not larger, than the differences between their means. A probability density plot of both sets of polls, *chart 1*, demonstrates this significant overlap in pro-independence voting intention found by polls that do and do not weight by 2014 vote.



To ensure that the differences in mean pro-independence voting intention are likely to be 'real' - that is, not the result of statistical noise or 'error' - we need to carry out statistical testing. The distribution of Yes voting intention for 2014 vote weighted polls looks close to a 'normal distribution', but the distribution for non-2014 vote weighted polls looks less so. This matters because the kind of test we use to determine whether the difference in mean pro-independence voting intention is likely to be 'real' depends on the normality, or otherwise, of these distributions.

To double-check, we can carry out Shapiro-Wilks tests for normality, which tests whether or not a distribution varies statistically significantly from a 'normal' distribution. In both the case of 2014 weighted polls ($p = 0.0696$) and non-2014 weighted polls ($p = 0.1876$) we can assume that the distribution of Yes voting intention does not vary significantly from the normal distribution.

We therefore use a one-tailed (as our hypothesis is directional) t-test to determine whether or not the mean pro-independence voting intentions for the two groups of polls differ statistically significantly - and they *do* ($t = 4.386$, $df = 45.98$, $p < 0.0001$).

So we can say that the mean pro-independence voting intention reported by polls that are weighted to reflect the 2014 Scottish independence referendum result is statistically significantly *lower* those that do not weight in this way, by 3.1 percentage points.

Regression Modelling

We know that there is a statistically significant difference of a few percentage points between pro-independence voting intention in Scottish polls, depending on whether they are weighted to reflect the outcome of the 2014 Scottish independence referendum, or not.

However, this is not sufficient to say that there is an effect of 2014 vote weighting on reported pro-independence voting intention. In the first instance, these measures are *correlative*, not *causal*. In the second, without accounting for other factors that may affect reported voting intention we cannot know if the association we have found is a result of 2014 vote weighting, some other set of factors that are associated with both 2014 vote weighting and reported pro-independence voting intention, or a mix of the two.

We therefore need to conduct some multivariate modelling. The models presented below use Ordinary Least Squares (OLS), a basic form of regression modelling. It is imperfect, but proves sufficient to reach some conclusions.

I have included the following predictors:

1. Poll is weighted to reflect the result of the 2014 Scottish independence referendum.
2. Poll is weighted to reflect the constituency vote in the *most recent* Scottish Parliament election.
3. Poll is weighted to reflect the Scottish vote in the *most recent* UK Parliament election.
4. Poll was conducted via an online panel.
5. Poll was conducted face-to-face.
6. Poll was conducted using the UK General Election franchise, that is adults age 18 and over.

All of these variables are 'dummy' variables. For predictors 4 and 5, their effects are in comparison to polls conducted by telephone. For predictor 6, its effect is in comparison to polls that use the Scottish election franchise - adults age 16 and over.

I have chosen these for two reasons. Firstly, we have good reasons to expect that they might have an effect on reported pro-independence voting intention:

- Predictor 1 - as we have seen, we know that there is an association between weighting to 2014 vote and reported pro-independence voting intention.
- Predictors 2 and 3 - it is generally accepted in political polling that weighting to past vote can affect the outcome of a poll, and various weighting strategies are employed by different research agencies.
- Predictors 4 & 5 - how poll participants are recruited and interviewed can affect the outcome of a poll for various reasons: for example an online panel may be biased in a particular direction if those who sign up to it are skewed in a given political direction, and may include more politically engaged people than the population; face-to-face polling may suffer from social desirability bias, and particpants may be less willing to disclose their true beliefs to a human interviewer than they would in an online poll.
- Predictor 6 - we know that younger voters are more likely to support Scottish independence, so excluding them from polls may lead to lower reported pro-independence voting intention.

Secondly, all of these predictors are relatively easily identifiable from publicly available sources. There are many other aspects of polling methodology that might affect outcomes, but these are either difficult to identify without working on a given poll (e.g. question ordering, online panel quotas) or difficult to quantify (e.g. interviewing style), or both (e.g. panel blend).

Moreover, as most Scottish independence polls are conducted by a handful of research agencies with largely consistent methodologies, it is likely the case that adding more and more predictors of this kind would be unproductive as the predictors we already have may well control for those we do not.

Models 1 - 4

Beginning with 2014 vote weighting, predictors are gradually added to build regression models 1 - 4 in *table 4*.

Model 1, which includes only 2014 vote weighting, shows a statistically significant relationship between 2014 vote weighting and reported pro-independence voting intention. Based on this bivariate model, we would expect a poll that weights by 2014 vote to find a level of pro-independence voting intention 3.06 points lower than a poll that does not.

However, this model has a low *adjusted R^2* value - it explains just 11.9% of variation in the reported Yes vote. It's not very good at explaining the variation in pro-independence voting intention between polls.

As we add more predictor variables, the *adjusted R^2* improves incrementally. By *model 4*, it has risen to 21.9% - better, but not great.

2014 vote weight remains statistically significant through all of these models, but by *model 4* its effect size - the degree to which we expect it to affect reported pro-independence voting intention - has fallen slightly, and we expect that a poll that weights by 2014 vote would report a pro-independence voting intention 2.94 points lower than a poll that does not.

And it appears that some of our other predictors are also statistically significantly associated with reported pro-independence voting intention. We would expect a poll conducted face-to-face to find pro-independence voting intention 6.69 points lower than one conducted by telephone, and a poll excluding 16- and 17-year-olds to find pro-independence voting intention 1.51pts lower than one that does not.

Table 4: Regression Models

	Yes Voting Intention			
	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.469***	0.479***	0.484***	0.485***
	(0.005)	(0.006)	(0.007)	(0.007)
Weighted by 2014 Indyref Vote	−0.031***	−0.023***	−0.028***	−0.029***
	(0.006)	(0.006)	(0.007)	(0.007)
Weighted by Holyrood Constituency Vote		−0.008	0.002	−0.0003
		(0.006)	(0.007)	(0.007)
Weighted by Westminster Vote		−0.019***	−0.010	−0.011*
		(0.006)	(0.007)	(0.007)
Online Panel Methodology			−0.011	−0.006
			(0.010)	(0.010)
Face-to-face Methodology			−0.069***	−0.067***
			(0.022)	(0.022)
18+ Sample				−0.015***
				(0.005)
N	204	204	204	204
R^2	0.124	0.171	0.210	0.242
Adjusted R^2	0.119	0.159	0.190	0.219

*p < .1; **p < .05; ***p < .01

So we have a model, albeit an admittedly rather poor model if what we want to do is predict the pro-independence voting intention a given poll will report. This is to be expected, as we have excluded an enormous number of variables that we might expect to affect not just *reported* voting intention, but *actual* support for independence in the population - like the state of the economy, evaluations of the Scottish and UK Governments, or voters' identities, values, and other attitudes.

Luckily, that was not our aim. Our aim was to determine whether or not we can say, with confidence, that weighting data to reflect the result of the 2014 Scottish independence referendum makes a meaningful difference to reported pro-independence voting intention. We seem to have demonstrated that it does.

But we now need to check that our model is robust. A number of assumptions underpin OLS, and we need to make sure that our model satisfies those assumptions. We therefore carry out a number of robustness checks.

In statistics parlance, we want our model to be 'BLUE' - Best Linear Unbiased Estimator. In other words, we want our model's predictions for each poll to be linear, to minimise error, and to be unbiased. We also assume that our error terms (a measure of the difference between what our model predicts and what our dataset actually tells us) are normally distributed, that there is no multicollinearity between our predictors (that our predictors do not predict for one another), and that we do not have any data points that are influential outliers (which can skew our model).

I will not be going into the detail of all of the tests conducted to check whether our model satisfies the

assumptions underpinning OLS, though the code for doing so is included in the R Markdown version of this document.

Having run these tests, we find that there is no multicollinearity between our predictors, and our error terms are normally distributed.

However, our model is incorrectly specified - perhaps because of omitted variable bias, which is to say that there is structure in the unexplained variance between our model's estimated pro-independence voting intention for each poll, and the pro-independence voting intention reported in that poll.

It also suffers from heteroscedasticity - our error terms are not independent of our predictions for each poll, they are related. Furthermore, our model features some autocorrelation - at least some data points are related to previous data points. We will return to these problems, which are fixable.

The fourth problem our model faces is that it features influential outliers. As it turns out, both of these outliers are TNS polls - the only polls in the dataset conducted face-to-face. There is a third TNS outlier which is not infuential.

I chose to remove these from the dataset and rerun the model. No Scottish independence polls are conducted face-to-face anymore, and TNS is no longer around as a pollster of Scottish independence voting intention, so their usefulness in this analysis is potentially limited. Given their detrimental impact on the model, it makes sense to exclude them from the analysis.

Model 5

With the TNS polls removed, our model - now *model 5* - still finds a statistically significant relationship between 2014 vote weighting and reported pro-independence voting intention. However, the effect size is smaller - we'd now expect a poll that weights to the 2014 result to report pro-independence voting intention 2.5pts lower than a poll that doesn't.

We would also expect a poll that weights by the Scottish vote in the most recent UK Parliament election to report pro-independence voting intention figures 1.31pts lower than a poll that does not, and one that excludes 16- and 17-year-olds to find figures 1.5pts lower.

Table 5: Regression Model 5

	Yes Voting Intention Model 5
(Intercept)	0.486^{***}
	(0.007)
Weighted by 2014 Indyref Vote	-0.025^{***}
	(0.007)
Weighted by Holyrood Constituency Vote	-0.003
	(0.006)
Weighted by Westminster Vote	-0.013^{**}
	(0.006)
Online Panel Methodology	-0.008
	(0.009)
18+ Sample	-0.015^{***}
	(0.005)
N	201
R^2	0.232
Adjusted R^2	0.212

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

We now re-run our robustness checks. While we still have some outliers, and some high-leverage data points, none are influential enough to be concerned about - and they certainly do not offer a justification for removing them from the dataset.

However, we continue to have issues with our model specification and heteroscedasticity. At this point, it is worth pointing out that - to an extent - our model deals with time-series data. We typically use OLS regression to deal with cross-sectional data (polls themselves are a classic example - data collected at one point in time, as a snapshot). But our data points, each poll, occur in a linear series over time. That may at least partial account for the unexplained structure in our data.

That we continue to also have an autocorrelation problem suggests that may be the right interpretation. So, the next step is to try to account for that.

Model 6

Typically, we deal with autocorrelation in OLS by introducing a lag variable. There are, of course, other and better methods for modelling time-series data to help us understand why a given variable changes over time. But, again, that is not the aim here. So we'll settle for introducing a lag variable in *model 6*.

Lagged Yes Vote Intention for each poll is equal to the pro-independence voting intention reported in the most recent previous poll. There are, of course, alternative lag variables - for example, the poll that came second-to-most recently, or third-to-most recently. Or, the poll immediately before the given poll but within the same research agency's polling series. For this reason, in addition to *model 6* we should construct models with alternative lag variables.

Alternative lagged models find largely the same relationships as *model 6*, but are increasingly plagued by autocorrelation the further from the given poll that the lag variable is drawn. Polls reflect other polls around them more than polls further in the past, even within a research agency's own series. The code for these models and their robustness checks is included at the end of the R Markdown version of this document.

Model 6 is our conclusive model. It suffers from none of the issues with our previous models, is BLUE, and satisfies all of the assumptions underpinning OLS regression.

We find that three of our predictors have statistically significant relationships with reported pro-independence voting intention. We would expect a poll that weights by 2014 vote to report pro-independence voting intention that is 1.86pts lower than a poll that does not; a poll that weights to the Scottish vote in the most recent UK Parliament election would find figures 1.35pts lower than a poll that does now, and a poll that excludes 16- and 17-year-olds would find figures 1.46pts lower than a poll that does not.

Our lag variable is also statistically significant, but does not require (and should not recieve!) interpretation similar to our other predictors. The *adjusted R^2* is higher than our other models - *model 6* explains 31% of variation in reported pro-independence voting intention - but this is entirely down to our lag variable.

Table 6: Regression Model 6

	Yes Voting Intention Model 6
(Intercept)	0.346^{***}
	(0.027)
Weighted by 2014 Indyref Vote	-0.019^{***}
	(0.006)
Weighted by Holyrood Constituency Vote	-0.007
	(0.006)
Weighted by Westminster Vote	-0.013^{**}
	(0.006)
Online Panel Methodology	-0.016^{*}
	(0.009)
18+ Sample	-0.015^{***}
	(0.005)
Lagged Yes Vote Intention	0.321^{***}
	(0.059)
N	200
R^2	0.331
Adjusted R^2	0.310

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Conclusions

Whether or not a poll weights by 2014 vote has a statistically significant, but quite small association with its reported pro-independence voting intention once we control for other broad methodological choices. Whether this effect size would continue to shrink, or disappear entirely, should we control for more factors is a question of interest.

This does not necessarily mean that it is necessarily wrong for a research agency to weight by 2014 vote, or not - that discussion is beyond the scope of this write-up, but there are arguments both ways. Indeed, it is valuable to have a plurality of methodological approaches to measuring public opinion on such a sensitive topic.

About the author

Mark McGeoghegan is a doctoral researcher at the University of Glasgow, specialising in the study of secessionist movements, political contention, and political violence. He also has a background as a public opinion researcher, and writes about Scottish politics and the politics of secessionism in several outlets.