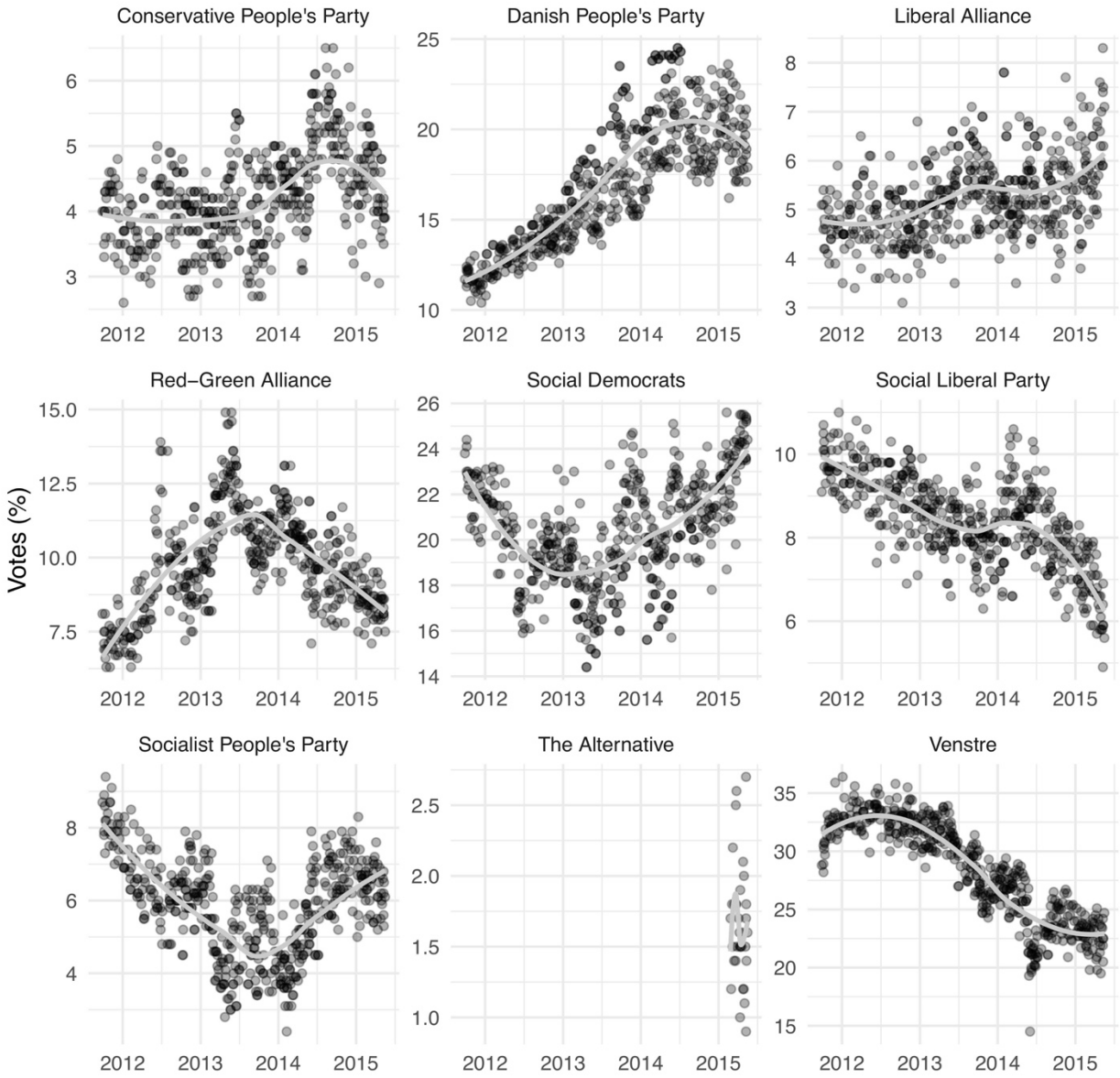
Supplementary Information:

"Transforming Stability into Change: How the Media Select and Report Opinion Polls".

1 Context and descriptive statistics

In the aftermath of the Danish general election in 2011, the Social Democrats, the Social Liberal Party and the Socialist People's Party formed a three-party coalition government with the support of the Red-Green Alliance. Figure A1.1 shows the support for nine political parties measured throughout this period.

Figure A1.1: Support for the parties, 2011-2015



In the period, there was a new entrant in Danish politics, The Alternative. The polling firms disagreed on the level of support for this party with some polls having them above the electoral threshold of 2% and other below, although a majority of these poll numbers were not statistically significant from each other. Given that this party has much fewer measurements as a new party, we do not include them in the main analysis.

The three government parties lost support during the first two years in office, partially due to a series of unpopular reforms and broken pledges. Some of the voters switched to parties within the red block, i.e. went to the Red-Green Alliance, whereas others went to the blue parties such as the Danish People's Party and Venstre. Towards the end of this period, the Social Democrats regained support in the public and Venstre lost a substantial number of votes and polled below 25% in a majority of the polls towards the 2015 general election. However, most importantly, none of these changes happened from one single poll to the next, and the year controls included in the reported models take any potential effects of these dynamics into account.

Table A1.1 shows descriptive statistics on the polls for each polling firm, including the number of polls, the average number of mentions in the coverage and the average volatility.

Table A1.1: Detailed descriptive statistics

Firm	Total polls	Earliest	Latest	Average mentions	Average volatility	Average significant changes	Days btw. polls
Epinion	39	2011-11-03	2015-05-12	6	3.105	0.18	34
Gallup	45	2011-10-06	2015-05-07	10	3.511	0.43	30
Greens	42	2011-10-06	2015-04-29	6	4.543	0.66	32
Megafon	49	2011-10-06	2015-04-30	22	4.224	0.35	27
Rambøll	75	2011-10-13	2014-09-15	4	4.242	0.59	14
Voxmeter	144	2011-09-28	2015-05-17	9	2.353	0.03	9
Wilke	18	2013-11-10	2015-05-10	7	3.732	0.29	32
YouGov	75	2011-10-10	2015-05-11	5	3.193	0.03	18

As noted in the main text, two polls (Rambøll, 2013-10-21 and Greens, 2014-06-08) were outliers in terms of volatility. In the Rambøll poll, the Social Democrats lost 5.4 percentage points and the Danish People's Party gained 4.7 percentage points. This was not a trend followed by other polling firms and it was an outlier poll for this firm as well. In the Greens poll, the Social Democrats gained 8.2 percentage points and Venstre lost 7 percentage points. Here, no other polling firms showed similar trends.

Last, in order to look into the context in which polls were selected and covered in more detail, we first examined the five most and five least volatile polls and whether they were selected or not. Second, we looked into the coverage of polls quoting sources and mentioning uncertainty for polls with low and high volatility.

For the polls covered, the highest volatility is 12.15% (as in total change in vote share estimate, Greens 2014-06-08), whereas for those not covered this value is 7.6% (Greens, 2013-10-25). Overall, we found no outliers in extreme polls not being covered, as all polls with the most extreme volatility were indeed covered. The most extreme polls covered and not-covered come from two polling firms, Rambøll and Greens. In other words, the most volatile non-covered and covered polls are from the same firms, confirming that our main selection results are not explained by systematic differences in whether a poll is picked up conditional on the polling firm.

For the polls with the smallest volatility covered and not covered, we see much less variation. This is explained by the fact that many polls show little variation. For lowest volatility polls we see an over-representation of polls from Voxmeter. While they are a large part of our population, the consistent weekly frequency of polls from them leads to only minor differences for a lot of the polls, some of which gets covered. One of the polls getting coverage is a poll with the story that a party, Venstre, gets "exactly 33.3 percent of the votes" (Voxmeter, 2011-11-27). Interestingly, the polls with the lowest volatility getting covered are framed as no change (e.g. Voxmeter, 2014-11-16 and 2014-11-23) or in relation to specific events that makes the support for the parties relevant, e.g. op-eds (Voxmeter, 2013-08-18) or the European Parliament election (Epinion, 2014-04-21).

For the polls with high volatility, we see that experts comment on the changes. The poll with the most coverage, Megafon (2012-05-31), resulted in articles where expert comments can be summarized as the changes were unique and that it had implications for the next parliamentary elections, with politicians commenting on the changes in the same way. A similar example is the poll from Rambøll (2013-10-10), that also resulted in pundits and politicians reacting to the poll due to the changes. Accordingly, we see that for the polls where there is a high degree of volatility, sources are included to comment on the changes, and in particular the causes of these changes and their potential consequences.

For the polls with the low volatility, once they get covered and there are reactions from sources, the comments vary from being about the support for a *specific* political party or potential future implications of these changes. In other words, there are cases when sources are included in low volatility polls, but they are less frequent and with no systematic pattern in what they comment on, in contrast to the overwhelming change focus in high volatility poll coverage.

2 Models of selection: additional details

Here we centralized all additional models, alternative specifications, and detailed considerations related to our argument about how change is associated with more frequent selection.

2.1 Alternative measures and specifications

In Tables A2.1 and A2.2 we report three models in each table, the only difference being that in the first table we have the robustness checks for the specification where we look at each poll as unit of analysis and have one overall media article count, whereas the second table reports models in which each poll has eleven (outlet) entries in terms of counts, with outlet × polling firm grouping and a hierarchical setup.

Table A2.1: Alternative models: negative binomial models of overall article count

	(1) Extreme values excluded	(2) Not significant only	(3) Maximum change
Intercept	2.72***	2.79***	2.71***
	(0.23)	(0.24)	(0.24)
Change (2 SD)	0.45**	0.57***	0.49***
	(0.14)	(0.17)	(0.14)
Any significant change (= 1)	−0.07		0.04
	(0.17)		(0.18)
Days since last poll (2 SD)	0.03	0.02	0.08
	(0.11)	(0.12)	(0.11)
Election campaign (= 1)	0.98***	1.29***	1.09***
	(0.28)	(0.30)	(0.28)
2012	−0.68**	−0.73**	−0.59*
	(0.25)	(0.26)	(0.26)
2013	−0.71**	−0.72**	−0.71**
	(0.25)	(0.26)	(0.26)
2014	−0.94***	−1.04***	−0.92***
	(0.25)	(0.26)	(0.26)
2015	−0.54*	−0.64*	−0.55
	(0.28)	(0.29)	(0.28)
AIC	2931.59	2401.98	3014.04
LogLik	−1455.80	−1191.99	−1497.02
N	473	392	479

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Year = 2011, baseline.

Model (1) in both scenarios lists results when we exclude the extremely high volatility polls (above 10%) identified and discussed in the paper, but also four polls that overall had more than 50 mentions. In the hierarchical setup the per outlet article number is lower, hence there we only

exclude the high volatility polls. Model (2) in both cases keeps only those polls that had no statistically significant changes compared to their previous counterparts, whereas finally, Model (3) substitutes volatility as a measure of change with the maximum change registered by a party for each poll.

Table A2.2: Alternative models: hierarchical negative binomial models of article count

	(1) Extreme values excluded	(2) Not significant only	(3) Maximum change
Intercept	−0.15	−0.11	−0.20
	(0.15)	(0.16)	(0.15)
Change (2 SD)	0.53***	0.51***	0.46***
	(0.06)	(0.06)	(0.06)
Any significant change (= 1)	0.14		0.12
	(0.07)		(0.08)
Days since last poll (2 SD)	−0.12**	−0.05	−0.09
	(0.05)	(0.05)	(0.05)
Election campaign (= 1)	1.06***	1.25***	1.03***
	(0.11)	(0.11)	(0.11)
2012	−0.68***	−0.85***	−0.60***
	(0.11)	(0.11)	(0.11)
2013	−0.71***	−0.77***	−0.66***
	(0.11)	(0.11)	(0.11)
2014	−1.09***	−1.20***	−1.06***
	(0.11)	(0.11)	(0.11)
2015	−0.65***	−0.72***	−0.60***
	(0.12)	(0.12)	(0.12)
AIC	10622.59	8390.63	10677.16
LogLik	−5300.30	−4185.31	−5327.58
N	5247	4312	5269
Dyads (outlet × company)	88	88	88
Var (Intercept)	1.22	1.26	1.22

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Year = 2011, baseline.

From an empirical point of view, as highlighted in the main text, these results bring further evidence that our main findings are not contingent on couple of extreme observations (1) or on the choice of change measurement (3). From a substantive point of view, we also showed that change matters even when the change is "illusionary" (2), as in statistical uncertainty related to polling estimates would suggest stability.

2.2 Media outlet count

To further our understanding of amplification beyond the number of articles, we also regressed the number of different outlets out of the 11 total outlets that report on a particular poll on the

predictors employed in the paper. In Table A2.3 we report the results from two models: in the first column all polls are included (including those with no reporting, i.e. 0 outlet count), whereas in the second column we subset our data to only those polls that received at least one mention. In both cases, we find evidence for amplification of reporting of change through diversification of the outlets reporting: polls indicating more change will be picked up by more different outlets.

Table A2.3: Model results: negative binomial model of outlet count

	Outlet count	Outlet count (one or more reports)
Intercept	1.84***	1.80***
	(0.17)	(0.12)
Change (2 SD)	0.38***	0.19*
	(0.10)	(0.08)
Any significant change (= 1)	−0.09	−0.01
	(0.13)	(0.10)
Days since last poll (2 SD)	−0.00	−0.07
	(0.08)	(0.06)
Election campaign (=1)	0.47*	0.35*
	(0.20)	(0.15)
2012	−0.58**	−0.41**
	(0.18)	(0.13)
2013	−0.56**	−0.32*
	(0.18)	(0.13)
2014	−0.76***	−0.48***
	(0.18)	(0.13)
2015	−0.46*	−0.35*
	(0.20)	(0.15)
AIC	2248.77	1874.26
LogLik	−1114.39	−927.13
N	479	404

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Year = 2011, baseline.

2.3 Hierarchical model extension

As referenced in the main text, we also fitted a model where the effect of volatility is let to vary across dyads, with correlation across varying effects fixed estimated. Model fit comparison

(Table A2.4) indicates that the effect of change does not vary across the outlet and polling firm combinations.

An alternative way to specify this model would be to use polls as the grouping level (or, level-2) and have a uniform 11 observations within each group, with volatility and other poll related quantities treated as level-2 predictors. However, this would mean that for all polls that were not reported, we would have no within-group variation, i.e. all outcome values would be 0. While this is not a fundamental issue, it would also mean that the only level-1 predictor is whether the outlet and the polling firm were partners. Letting this effect vary across polls and potentially adding a model to this variation with volatility being a predictor (cross-level interaction), would enable us to discuss whether the positive effect of partnership varies as a function of volatility. We can reach similar substantive conclusions regarding the effect of partnership as in our main approach, however, that relationship is not the core quantity or predictor of interest here.

Table A2.4: Model fit comparison: varying effect of change (across outlet × firm dyads)

	Df	AIC	BIC	deviance	Chisq	Chi Df	p-value
Original model	12	10623	10701	10598.62			
Varying slope of change	14	10626	10718	10597.93	0.69	2	0.7090

3 Reporting: additional details

3.1 Title coding task

The title coding was carried out in two steps, i.e. two questions:

(1) does it contain any mention of a party or party-block support (yes or no), if *yes*:
(2) what kind of support interpretation is given, with the options: (1) close race, (2) status quo, standstill, (3) party/block names is losing votes, (4) party/block names is winning votes, (5) one party/block is winning, another losing.

In the main text, we treat answers 2.3, 2.4, and 2.5 as (1 – there is change mention in the title), all the rest of the options, including no support mention (no for question 1) as 0. This is to provide a conservative test where we might underestimate the actual focus on change in the reporting. We have re-run our analysis using an alternative coding where close race is also coded as change. Inter-coder reliability is unchanged, the supervised machine learning results presented in the main text are slightly better as "close race" tends to mention parties or blocks as does change coverage, but all substantive results are the same. In other words, decisions related to how close race should be treated are not influential for our results.

3.2 Classifier summary

Once training the classifiers, for each particular categorization task, we can summarize which features (uni- or bi-grams build from stems) carry importance in assuring accurate classification. The terms listed in Table A3.1.1 and Table A3.1.2 (translated) are the top 40 most important terms for each of the three main classifiers (usually 110-130 features with non-0 importance). Again, these are not "directional" *per se* (although can be added based on what class of documents they appear more often). Instead, they indicate that if the texts contain these features, the classifier will do better in differentiating between the classes. While we do not list the exact gains in terms of prediction error reduction associated each feature, it is worth noting that the top 5-10 range carries the meaningful weight.

Table A3.1.1: Importance ranked terms

	Change in title	Uncertainty	Quote		Change in title	Uncertainty	Quote
1	måling	usikkerh	siger	21	i_ny	måske	ifølg
2	meningsmål	tokenanynumn_procentpoint	sagd	22	valget	fire	op_til
3	blok	procent	ritzau	23	røde	inden	målinger
4	sender	statistisk	politisk_ordfører	24	tokenanynumn	ritzau	kan
5	mellem	tokenanynumn	så	25	chokmål	blandt_andet	procent
6	nedtur	er_tokenanynumn	politisk	26	radikal	ved	udvalgt
7	ny	derfor	komment	27	tager	tidliger	tilbagegang
8	katastrofemål	repræsentativt	partiet	28	frem	gør	største
9	går	står	tror	29	historisk	politik	i_stedet
10	ved	foretaget	lyder	30	dag	opbakn_fra	lige
11	vælgern_stemm	den_tokenanynumn	valgforsk	31	store	over	bruge
12	politisk	person	gik	32	stormer	tilbagegang	til_tokenanynumn
13	får	vore	ordfører	33	mandat	i_folketinget	blandt
14	større	nogensind	universitet	34	fremgang	procent_af	politik
15	flertal	i_tokenanynumn	fordel	35	siden	senest	stor
16	spærregrænsen	meget	rød	36	lige	niveau	statsministeren
17	tilbag	godt	og_så	37	største	stadig	på_tokenanynumn
18	analys	langt	uger	38	vælgere	kommer	ting
19	giver	vælgern	gå	39	vælgern	svarer	står
20	regeringen	partiet	dansk	40	fast	landet	danskern

Table A3.1.2: Importance ranked terms (translated)

	Change in title	Uncertainty	Quote		Change in title	Uncertainty	Quote
1	poll	uncertain	says	21	in_new	maybe	according
2	opinion poll	tokenanynumn_percentagepoints	said	22	election	four	up_to
3	bloc	percentage	ritzau	23	red	before	polls
4	sender	statistical	political spokesman	24	tokenanynumn	ritzau	can
5	between	tokenanynumn	saw	25	shock	among_other	percentage
6	downturn	is_tokenanynumn	political	26	radical	by	selected
7	new	therefore	comment	27	takes	former	decline
8	disaster poll	representative	party	28	forward	do	largest
9	going	stands	believes	29	historical	political	instead
10	by	conducted	says	30	Day	support_from	equal
11	voter_vote	the_tokenanynumn	researcher	31	big	over	use
12	political	person	went	32	storms	decline	to_tokenanynumn
13	gets	our	spokesman	33	mandate	in_parliament	among
14	larger	ever	university	34	progress	percentage_of	political
15	majority	in_tokenanynumn	advantage	35	since	latest	large
16	threshold	very	red	36	equal	level	primeminister
17	back	good	and then	37	largest	still	on_tokenanynumn
18	analys	Long	weeks	38	voters	coming	thing
19	provides	voters	walk	39	voter	similar	stands
20	government	party	danish	40	firm	landed	danes

3.3 From title change to change in text

As highlighted in the main text, there are operational and theoretical reasons to focus on the title text rather than the full text of the article when assessing the change mentions in the reporting. Two additional empirical considerations underline that there is no systematic bias in favour of confirming our second hypothesis. In the results in the main text we find no evidence for a relationship between change in titles and change in polls, but we do find that change in title is a frequent part of the reporting, i.e. even small volatility polls are presented as change once they get through the selection change.

In our first approach, we randomly selected 40 articles, equally split between labelled as title having change or no change. An additional coder blind to the selection and goal coded whether change was mentioned in the content of each of the articles, based on reading the full text. Next, in a separate file (differently ordered), the coder was asked to do the same based on the title text alone, mimicking the task in our main sections of the paper. According to this coding step, the proportion of change mentions coded based on titles was 0.4 (compared to 0.5 in the data), but the proportion was 0.775 based on the coding of the full text. This indicates that through the coding of titles alone, we are likely underestimating the change reporting compared to what we were to get based on coding the full text. When we subset our extra coding dataset, we see that these numbers are 0.75 for articles where original title coding was done by other coders and 0.8 for those where the labelling is the result of the machine learning.

To reiterate, the main aim here is not to explicitly think of correspondence between coding based on titles *vs* full text, rather to see in which direction the differences appear. In this regard, while a limited exercise, 31 out of the 40 articles were coded to have change mentions *based on the full text,* and in our data 17 of these are labelled as not being about change *based on the titles*. Overall, we saw good human coding and machine learning performance based on titles and those are used for our main analysis; if we were to think of reporting features based on full article texts, we should expect that change coverage is even more often mentioned.

In our second approach, we modify the prediction steps of our machine learning approach to further substantiate that full text based change coding could only strengthen our claim that change is ubiquitous in the media reporting about opinion polls. As a first step, similar to our main analysis, we trained our classifier using the title texts. We trained two classifiers, one with no

reduction in terms of number for features (9,990 features in total) and one that yielded the best performance (174 features).

For the random 20% subset of our data (*test set, not included in training)*, we used the trained classifier to predict labels of change or no change. To do this, we used the document-feature matrix created according to the same rules of the full texts. To be more precise, we used the features that were present in the training set (titles) and also in the full texts. However, these features came with different frequencies. When we used the full text on the *test set* prediction, the average proportion of change was 0.877 (using all features, no sparsity reduction) and 0.872 (reduced sparsity, features present in at least 24 titles) respectively. On the same test sets, if we use the title based document-feature matrix and the same classifier, the proportion change was 0.64 and 0.63. Thus, the first take-away, consistent with the human coding exercise, is that full texts would only indicate higher change reporting. If we conduct any transformation (such as *tf-idf*) to account for the length of full texts, these numbers will be even higher, above 0.9. In terms of interpretation, this simply indicates that full texts contain words associated with change derived from the titles, and proportionally they carry an important weight. It does not necessarily mean that 80% of the articles are only about change, but they do make enough change references (more than one) to be regarded as change reporting.

As human coders worked with title texts for this task and the coding is based on that, accuracy on the test set using title texts is much higher than that using full text (0.839 *vs* 0.645). Furthermore, likely both change and stability vocabulary is larger for full texts. What is noteworthy is that out of 75 no change labels by the human coders in the test set, the prediction based on titles mislabels 17 as being about change, but this is 65 for the case full text content. This final piece of information is to underscore that using titles to capture change reporting is unlikely to bias the findings upwards, i.e. to overestimate the amount of change reporting.

3.4 Detailed consideration of quote sources

To provide additional information on the analysis of quoted sources, we keep the original coding of quotes with minor reduction of complexity only. Specifically, we do not differentiate between red and blue block politicians (treated as Politician) and between university affiliated experts or political commentators (treated as Expert). We follow the same procedure as before for human coding and machine learning.

The inter-coder reliability and agreement for the 4-category[1] quote measure was in line with other numbers reported in the main text, i.e. very good: 90%/0.86 (coder1:coder2), 89%/0.84 (coder 1:coder 3), 92%/0.88 (coder 2:coder 3). For the supervised machine learning the 0.1 lower threshold was deemed best, resulting in a dfm with 422 features. A multinomial classifier with boosting was used, resulting in good performance given the difficulty of the task: 0.82 accuracy and confusion matrix reported in Table A3.2. F1 scores for each category in order were: 0.84, 0.86, 0.79, and 0.70. After labelling the full dataset, 32% had no quote, 37% quoted a politician, 19% an expert, and 12% of articles quoted both.

Table A3.2: Confusion matrix

	Observed			
	0	1	2	3
Predicted 0	52	7	3	0
1	7	70	1	1
2	2	3	34	1
3	2	3	9	20

We fitted four hierarchical models (with grouping at the poll level) for the 4-category quote split up. We do not fit a hierarchical multinomial model because its complexity and the cases of no-within poll variation (only one report per for a poll) create estimation difficulty. The results are reported in Table A3.3.

Two consistent findings emerge: as volatility increases, there is a sharp drop of "No quote" scenario compared to all other options together, and an increase in the probability of quoting a politician compared to all other options together. We also find that change is not associated with

[1] 0 = No quote, 1 = Politician quoted, 2 = Expert quoted, and 4 = Both quoted.

expert quoting. Finally, while quotes covering multiple types of sources are more likely with higher change polls, these differences are also not statistically significant.

Table A3.3: Hierarchical models of quote type

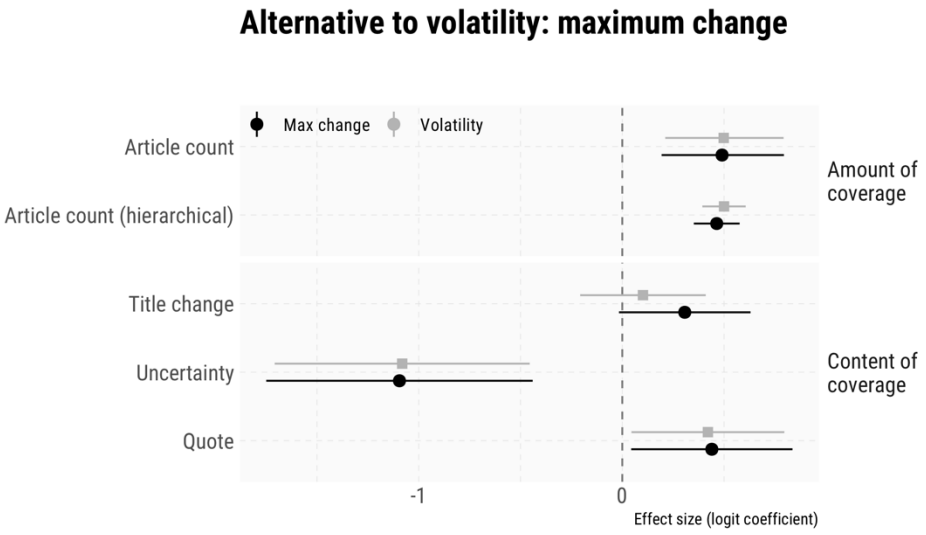
	No quote	Politician quote	Expert quote	Both quoted
Intercept	-1.47***	-0.23	-2.42***	-1.99***
	(0.28)	(0.27)	(0.35)	(0.38)
Change (2 SD)	-0.51**	0.44*	0.03	0.28
	(0.19)	(0.19)	(0.22)	(0.28)
Any significant change (= 1)	0.03	0.15	-0.41	0.29
	(0.22)	(0.22)	(0.26)	(0.31)
Days since last poll (2 SD)	-0.27	0.09	-0.08	0.55**
	(0.14)	(0.14)	(0.17)	(0.21)
Election campaign (= 1)	0.08	-0.03	0.28	0.13
	(0.32)	(0.34)	(0.37)	(0.49)
Partner (= 1)	-0.31**	-0.02	-0.10	0.68***
	(0.10)	(0.10)	(0.12)	(0.13)
2012	0.60*	-0.73*	1.04**	-0.68
	(0.30)	(0.30)	(0.37)	(0.42)
2013	0.43	-0.45	0.81*	-0.61
	(0.30)	(0.30)	(0.37)	(0.42)
2014	0.80**	-0.67*	0.86*	-1.09*
	(0.30)	(0.30)	(0.37)	(0.43)
2015	0.99**	-0.73*	1.15**	-1.94***
	(0.33)	(0.34)	(0.41)	(0.52)
AIC	4536.51	4770.07	3612.40	2616.25
Articles	3824	3824	3824	3824
Polls	402	402	402	402
Var (Intercept)	0.89	1.00	1.18	1.84

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

4 Alternative measure of change

We have used volatility between polls to measure change in our analysis as it fits with our case of multiparty competition. It accurately captures overall changes and can be extended to any other party system, both with more and with fewer parties regularly measured in the polls. However, there are alternative measures that would build on changes in the standing. Most certainly, this potential to measure overall changes through taking into account all party changes also makes it an unlikely candidate to being employed directly by journalists when evaluating the narrative potential and taking a decisions about selection.

Figure A4.1: The relationship between maximum change and previously studied outcomes



An accessible and intuitive heuristic for a change narrative is the greatest magnitude of change. Rather than summarizing it into one measure, for each poll, we looked at the maximum change, in absolute terms, a party registered compared to the previous poll from the same firm. This is directly apparent after looking at a poll in comparison to previous numbers, and allows for a more party centric coverage, i.e. the biggest winners and losers. We re-fitted all previous models, but instead of volatility, we used the maximum change (mean centred and standardized

with two standard deviations). As we followed the same steps, these results should be directly comparable between the two operationalizations.

As displayed in Figure A4.1, we see very strong consistency in our results. This is not surprising, as the correlation between volatility and maximum change is 0.89. It is reassuring, as it shows that the empirical support for our theoretical model is not contingent on the specific operationalization, and that simple heuristics that might be more fitting for journalistic decision-making models work equally well.