How News Media Affects Political Behavior

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF POLITICAL SCIENCE AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

NA RYEUNG WHANG AUGUST 2020

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

This dissertation is online at: http://purl.stanford.edu/fy882yn1499

Includes supplemental files:
1. Supplemental file for ETD druid:fy882yn1499 *(p2-appendix.pdf)*

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Justin Grimmer, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Andy Hall, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Shanto Iyengar

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Checking How Fact-checkers Check

August 28, 2020

Abstract

Fact-checking has gained prominence as a movement to revitalize truth-seeking ideals in journalism. While fact-checkers are often assumed to code facts accurately, few studies have formally assessed fact-checkers' overall performance. I evaluate the performance of two major fact-checkers, Fact Checker and Politifact, comparing their interrater reliability using a method that is regularly utilized across the social sciences. Surprisingly, only 1 in 10 statements is fact-checked by both fact-checkers. Among claims that both evaluate, fact-checkers perform fairly well on outright falsehoods or obvious truths. However, the agreement rate is much lower for statements in the more ambiguous scoring range (i.e., "Half True" or "Mostly False"). The results suggest that fact-checking is difficult, and challenging to validate. Fact-checkers rarely evaluate the same statement and disagree more often than one might suppose, particularly when when politicians craft language to be ambiguous. Politicians' strategic ambiguity may impede the fact-checking movement's goals, at least in some cases.

1 Introduction

Often described as the umpires of democracy, fact-checkers aim to keep false political claims out of the public discourse (Amazeen 2013). According to Graves (2016), fact-checking may inoculate readers against deceptive claims and inhibit political lying by making it more costly for political figures to distort the truth. But how easy is it for fact-checkers to do their job, in practice?

I evaluate the performance of two major online fact-checkers (Politifact and Fact Checker) by using a method that is regularly utilized across the social sciences for assessing interrater reliability among multiple coders. I show that fact-checkers rarely fact-check the same statements and disagree more often than one might suppose. I conclude that fact-checking is challenging to validate, partly because fact-checkers do not use directly comparable scales. Also, assessing the significance or the implications of inaccuracies in political claims may require subjective judgment, which often leads to discrepancies in ratings. I find that in many cases, these discrepancies can be explained by one of the following: 1) fact-checkers present the same set of counterexamples or evidence but have different views on the significance or the implications of them; 2) fact-checkers vary in the number of counterexamples or evidence in support of or against the statement in question.

Previous works have focused on the effects of fact-checking. Some suggest that fact-checking can serve as an "extensive and consistent monitoring [tool for] politicians" by discouraging them from promoting false or misleading claims (Nyhan and Reifler, 2015). Others have argued that fact-checking has had very little effect in changing candidate behavior (Froomkin, 2012; Gottfried et al., 2013). More specifically, candidates often ignore fact-checkers' critique by simply dismissing the fact-checking community as politically biased (Waldman, 2015). Donald Trump, for example, repeated many claims despite the negative ratings he had received for those statements. The Trump campaign, according to the *Washington Post* Fact Checker Glenn Kessler, "does not even bother to respond to fact-checking inquiries" (Kessler, 2016). Trump was only one among the many candidates the fact-checkers never heard back from when they asked for a clarification or correction for inaccurate statements.[1]

While the effects of fact-checking have been studied by numerous scholars, very few studies have formally assessed the overall performance of fact-checkers. In my evaluation, I focus on the two aspects fact-checkers claim make their work effective: 1) many political claims are fact-checked by multiple fact-checkers, thus allowing for a second pair of eyes and 2) fact-checkers tend to agree on the accuracy of a given political claim.

Based on extensive fieldwork and interviews conducted over the course of 5 years at major fact-checking outlets, Graves (2016) points out that unlike traditional news reporters who deliberately try not to repeat the news that has already been published, fact-checkers are less intent on generating a scoop or exclusive news content. According to Graves (2016), Angie Holan, Eugene Kiely, and Glenn Kessler, chief editors at each of the three major online fact-checking outlets, "watch [one another]'s sites and repeat [one another]'s fact-checks with [their] own process" and "get a little ruffled" when they reach different conclusions. Fact-checking has a lot to offer especially when politicians craft language to be ambiguous. Due to the difficulty of validating the truth of these subtle forms of deception, the public could benefit from having multiple fact-checkers independently evaluate politicians' misleading remarks. As Kessler argues, fact-checking can have the greatest impact when multiple fact-checkers reach the same conclusion on a given statement. Thus, it is important to analyze 1) whether political claims are in fact evaluated by multiple fact-checkers 2) if fact-checkers are able to reach the same conclusion on how accurate a given claim is.

To assess fact-checkers' performance, I have come up with an effective research design which differs from previous studies in the following ways:

First, my design examines an expansive set of political claims. Amazeen (2016), who was among the first to provide a formal assessment of fact-checkers' performance, focuses on fact-checks of political claims in campaign ads. Similarly, Marietta et al. (2015)'s findings are confined to a small number of topics. My sample, however, includes all statements made by candidates of the 2016 U.S. presidential election that were fact-checked by Politifact and Fact Checker.

Second, I use two different metrics to assess the performance of fact-checkers. Previous studies have focused exclusively on how often fact-checkers agree on a set of identical statements (Amazeen 2016), while neglecting to examine how often a political claim is evaluated by more than one fact-checker. In this paper, I evaluate both if a given claim is independently evaluated by multiple fact-checkers and if fact-checkers are able to reach the same conclusion on the accuracy or truthfulness of the claim.

Third, I assess the performance of fact-checkers based on their own ordinal scales rather than arbitrarily converting their scales into a binary or a ternary scale. As I show in Section 2, fact-checkers' consensus rates vary widely depending on the researcher's choice of conversion method. Moreover, the researcher's arbitrary truncation of the fact-checkers' original ordinal scale may obscure the nuances fact-checkers may have intended to convey through a more fine-grained, ordinal scale. Thus, I use the ordinal scale to offer the fact-checkers the fairest assessment possible.

My results show that while fact-checkers perform fairly well on outright falsehoods or obvious truths, the agreement rate is much lower for statements in the more ambiguous scoring range (i.e., "Half True" or "Mostly False"). Lack of consensus among fact-checkers may arise from the challenge of verifying the accuracy of political claims. Politicians tend to be quite vague (Shepsle, 1972), which makes it difficult for fact-checkers to evaluate claims in a clear and objective manner. Because fact-checkers rarely evaluate the same statement and disagree more often than one might suppose, fact-checking may fall short of holding politicians accountable for their words.

2 An Original Dataset on Fact-checks and Ratings

2.1 Data

I test two of the main factors that fact checkers suggest make their work effective: 1) large overlap and 2) high consensus rate, by analyzing fact-checks on 2016 presidential candidates' statements from September 2013 to November 8, 2016. September 2013 is chosen because it is the earliest date from which fact-checks have been systematically archived on two major fact-checking outlets, Politifact and Fact Checker. FactCheck.org, another major fact-checking outlet, is excluded because unlike Fact Checker or Politifact, FactCheck.org does not provide ratings for politicians' statements. Therefore, only Fact Checker and Politifact are used for a comparative analysis.

As shown in Figure 1, 1178 and 325 fact-checks were obtained from Politifact and Fact Checker, respectively.[2] If a statement was fact-checked only by either one of the two fact-checkers, it was labeled as Nonoverlap. Pairs of identical or similar statements that were fact-checked by both fact-checkers were categorized as either Overlap or Murky (See Appendix C and D for a complete list of Overlap and Murky statements). Overlap and Murky statements were hand-coded as follows: Because Fact Checker evaluated fewer statements than does Politifact, for each statement on Fact Checker, I examined Politifact's fact-checks for a corresponding candidate to find a statement that was similar (Murky) or identical (Overlap). Interrater Reliability Rate among two independent coders on 50 randomly selected fact-checks was 0.75, which is well within the 0.7-0.8 range, the standard of interrater reliability that academics commonly apply when evaluating hand-coded data (Barrett, 2001).

Distinctions between Murky and Overlap were made based on the title of the article which provides information on the statement being fact-checked, each fact-checker's evaluation objective for a given statement, and fact-checkers' explanation for their ratings. A pair of two identical statements was coded as Overlap. In some cases, however, the statements being fact-checked were almost identical except that 1) one version included an additional

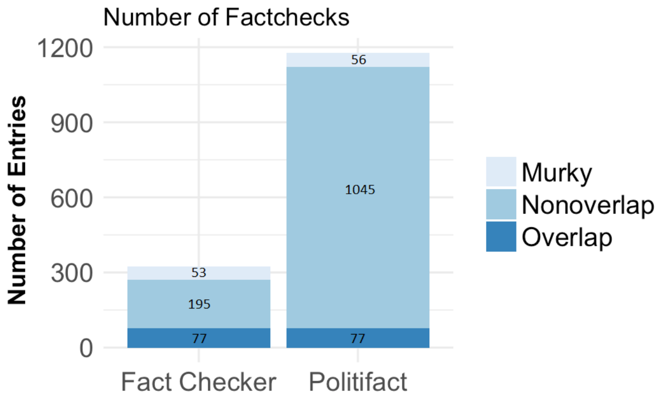


Figure 1 – Number of Fact-checks by Politifact vs. Fact Checker Among the 1178 statements fact-checked by Politifact, only 6.5 percent (77 statements) were also evaluated by Fact Checker. Of the 325 claims fact-checked by Fact Checker, only 23.7 percent were also assessed by Politifact.

phrase/sentence and/or 2) one fact-checker examined only parts of a claim whereas the other fact-checker evaluated the entire claim and/or 3) the key phrase was worded differently. These statements were coded as Murky. For example, “We have the highest murder rate in this country in 45 years” and “We have an increase in murder within our cities, the biggest in 45 years” were coded as Murky, because the former focused on the murder rate itself whereas the latter considered the size of an increase in murder. Yet, non-identical statements could still be coded as Overlap if different wordings did not affect fact-checkers’ evaluation or were not the key point of contention. For instance, the following pair of two non-identical statements – “58 percent of African American youth is unemployed” and “The unemployment rate for African-American youths is 59 percent” – were coded as Overlap, because the key point of contention for both fact-checkers was whether or not Trump exaggerated the figure, not whether the true unemployment rate was 58 or 59 percent.

2.2 Lack of Overlap in what Fact-checkers Evaluate

According to chief editors at the three major fact-checking websites, fact-checkers often repeat one another’s fact-checks and find it important that they reach the same conclusion on a given political claim (Graves 2016). Due to the difficulty of validating the accuracy

Table 1a. Proportion of Overlap
The proportion of overlap are 0.065 and 0.237 for Politifact and Fact Checker, respectively. When "murky" statements are included, the proportion of overlap rises to 0.113 for Politifact and to 0.400 for Fact Checker. Roughly, only 7 in every 100 statements in Politifact was also evaluated by Fact Checker.

	Overlap	Overlap + Murky
Politifact	0.065	0.113
Fact Checker	0.237	0.400

Table 1b. Consensus Rate
For statements that are fact-checked by both Politifact and Fact Checker, weighted Cohen's κ is 0.750 and the unweighted κ is 0.467. When "murky" statements are included, the intercoder reliability slightly decreases on both weighted and unweighted scales.

	Overlap	Overlap + Murky
Weighted	0.750	0.668
Unweighted	0.467	0.324

of political claims, the public could benefit from having multiple fact-checkers verify one another's ratings by independently evaluating a given claim. A high rate of overlap implies that a large number of political claims are being scrutinized by more than one fact-checker.

I measure the rate of overlap by computing the following: $\frac{\text{Number of Overlap Statements}}{\text{Total Number of Statements}}$ and $\frac{\text{Number of Overlap Statements + Number of Murky Statements}}{\text{Total Number of Statements}}$. Murky statements are included to credit fact-checkers for evaluating claims that are very similar (though not identical word-for-word). Among the 1503 fact-checks, there are 77 pairs of Overlap statements, 53 Murky statements for Fact Checker, and 56 Murky statements for Politifact[3] (See Figure 1). The proportion of overlap are 0.065 and 0.237 for Politifact[4] and Fact Checker, respectively. Roughly, 93 percent of statements that are fact-checked by Politifact are not evaluated by Fact Checker. Likewise, more than 75 percent of claims that are fact-checked by Fact Checker are not covered by Politifact. After including murky statements, the rate of overlap rises to 0.113 for Politifact, and 0.400 for Fact Checker. Even then, however, it appears that in the majority of cases, fact-checkers do not fact-check the same statement, and as a result, readers are not afforded a second pair of eyes for an accountability check.

2.3 Measuring Consensus among Fact-checkers

Fact Checker and Politifact each operate on a similar (although not directly comparable) rating system. Politifact uses a 6-point scale (True, Mostly True, Half True, Mostly False,

Figure 2 - Confusion Table for Overlap Statements

		Politifact Ratings				
		1	2	3	4	5
Fact Checker Ratings	1	4	0	0	0	0
	2	2	2	1	1	1
	3	0	0	4	4	2
	4	0	0	5	10	6
	5	0	1	1	4	29

False, and Pants on Fire) while Fact Checker has a 5-point scale (Geppetto Checkmark, 1 Pinocchio, 2 Pinocchios, 3 Pinocchios, and 4 Pinocchios). To compute any correlations, I had to make Politifact's 6-point scale comparable to Fact Checker's 5-point scale. Following Kessler (Chief Editor at Fact Checker)'s own interpretation[5] of how Fact Checker's scale compares to that of Politifact's, I group Geppetto Checkmark with True, 1 Pinocchio with Mostly True, 2 Pinocchios with Half True, and 3 Pinocchios with Mostly False, and both False and Pants on Fire with 4 Pinocchios.

As mentioned in the introduction, I assess the performance of fact-checkers using their own ordinal scale rather than a truncated binary or ternary scale, because fact-checkers' interrater reliability varies depending on the researcher's choice of conversion method. For example, when Amazeen (2016)'s coding scheme (i.e., 0 if True or Geppetto Checkmark and 1 if Mostly True, Half True, Mostly False, or False) is applied, the interrater reliability between Politifact and Fact Checker is very high (0.787). However, using a ternary design in which I classify True (Geppetto Checkmark) as 1, Mostly True, Half True, Mostly False as 2, and False, Pants on Fire as 3, I obtain a much lower interrater reliability score of 0.471 (See Appendix B for interrater reliability scores for various possible binary and ternary designs).

Cohen's κ coefficient was used to compute the interrater reliability between Politifact and Fact Checker. Cohen's κ is widely utilized in social sciences to measure the rate of agreement among multiple coders. This measure is considered more robust than the raw agreement rate, which, according to Cohen, is an inflated index because it fails to exclude agreements that happen by chance (Barrett, 2001).

The weighted κ, whose penalty increases at a higher rate as the size of disagreement becomes larger, is 0.750, which meets the typical threshold considered acceptable for evaluating hand-coded data. I also compute an unweighted κ because fact-checkers do not specify the distance between each category on their scoring scales. For example, "Mostly True" is two scales apart from "Mostly False" and likewise, "Half True" is two scales down from "True". However, the distance between "Half True" and "True" may not be equal to the distance between "Mostly True" and "Mostly False". The unweighted κ is 0.467 – lower than typical thresholds for scientific coding.

As shown in Figure 2, Politifact and Fact Checker agreed on the ratings for 49 out of 77 Overlap statements. Among the 28 cases in which they disagree, the ratings vary by two or more scales only 21.4 percent of the time. About 78 percent of the disagreements are relatively minor (i.e., the ratings vary by only one scale). Also, fact-checkers perform fairly well when evaluating outright falsehoods or obvious truths. As the confusion table shows, no statement received both Geppetto Checkmark and Pants on Fire! (See Appendix A for Confusion Tables for Overlap/Murky sample). Note that the Murky sample is not used to calculate the consensus rate between Politifact and Fact Checker to ensure that they are not penalized for discrepancies in ratings caused by an addition of a word to an otherwise identical statement.

Yet, the agreement rate is much lower (14 agreements out of 23 cases) for statements in the more ambiguous scoring range (i.e., Half True (point 3) or Mostly False (point 4)). I find that in many cases, discrepancy in ratings often stems from differences in fact-checkers' subjective judgment of the *significance* of inaccuracies within a statement, rather than their disagreement on whether the statement itself is true or not. For example, both fact-checkers evaluated Jeb Bush's claim that "Florida led the nation in job creation". The two fact-checkers provided identical sets of rationale for why Bush's claim may be misleading: 1) Bush relies on raw job totals; 2) the year 1999 was omitted; 3) much of Florida's job gains were due to an increase in low-paying jobs. Fact Checker decided that Bush deserved 4

Pinocchios while Politifact concluded that these exact same fallacies were not nearly as egregious, rating the claim Half True.

Another source of discrepancies in ratings is differences in the number of counterexamples or evidence each fact-checker uses in support of or against the statement in question. For instance, Fact Checker gave 3 Pinocchios (roughly equivalent to Mostly False) to Rick Perry's claim that "In the last seven years of [his] tenure, Texas created 1.5 million new jobs." while Politifact rated the claim Mostly True. Upon carefully analyzing the fact-checkers' explanation, it seems that Fact Checker gave a higher dishonesty rating because Fact Checker found an additional fault in Perry's statement. In addition to offering the same set of evidence presented by Politifact (i.e., cherry-picked data sources), Fact Checker also pointed out that he had aggregated unemployment numbers in an incorrect manner. However, both Fact Checker and Politifact agreed that Perry's claim is not completely free of inaccuracies.

Oftentimes, Half True or Mostly False statements are subtle claims that politicians often use to be deceptive. Underneath what appears to be a "true" statement, politicians may attempt to mislead by engaging in logical fallacies such as cherry-picking a more favorable piece of evidence or using a straw man argument. Fact-checks of these subtle forms of deception are what readers could most benefit from; yet, this is also where fact-checking struggles the most. My findings imply that this is partly because determining the significance of logical fallacies or inaccuracies in a given context may require subjective judgment.

3 Content of Fact-checks

In this section, using a text classification model, I analyze the content of fact-checks and evaluate if fact-checkers tend to agree more or less often in certain topic areas. To analyze the content of fact-checks, I first gathered 1503 fact-check entries from Politifact and Fact Checker. Each entry consists of a direct quote of the statement being fact-checked and the fact-checker's evaluation of the statement. I created a document term matrix by first

Table 2 LDA-classified Topics and Keywords for all fact-checks The first column represents a topic number for each topic; the second column provides manually generated labels; and the third column corresponds to 10 most frequently occurring terms under each topic.

Topic	Keyword	Keys
1	Immigration	trump, immigr, peopl, clinton, donald, border, campaign, illeg, refuge, report
2	Social Policy	job, state, walker, school, work, student, worker, year, educ, wisconsin
3	National Security	state, unit, countri, obama, trade, iran, presid, deal, militari, rubio
4	Campaign	clinton, sander, vote, cruz, campaign, democrat, republican, support, senat, rubio
5	Healthcare	tax, billion, plan, health, budget, year, spend, million, fund, care
6	Clinton	law, state, clinton, gun, email, depart, court, case, feder, report
7	Economy	percent, rate, number, data, year, peopl, incom, state, bush, popul

stemming the words in the text, discarding stop words, and representing the words in unigrams (single words). Each row of the document term matrix represents a single fact-check entry and the column consists of the 1000 most commonly occurring unigrams. Each cell corresponds to the number of times a given unigram appears in a given entry.

Then, I applied latent Dirichlet allocation (LDA) to model the topics in the texts. I assume that the collection of fact-checks on Politifact and Fact Checker are driven by 7 topics, a number chosen after assessing the substantive fit within and among the clusters. LDA assumes that each word in a document is generated from a single topic. Since different words in a document may be generated from different topics, each document is represented as a mixture of different proportions of various underlying topics. I then assigned the topic with the maximum proportion to each document.

Using the output from LDA, topics are manually labeled. The first column of Table 2 represents a topic number for each topic; the second column provides manually generated labels; and the third column corresponds to the 10 most frequently occurring terms under each topic. Statements under Campaign directly pertain to election-related matters (e.g., performance of candidates in the polls). Topics such as Immigration, Healthcare, and the Clinton controversy were issues of high salience during the 2016 election.

In Figure 3, the 7 LDA-classified topics are on the x-axis and the level of attention dedicated to each topic is plotted on the y-axis. The proportion of Nonoverlap and Overlap/Murky statements within each topic area is represented in black and grey, respectively.

Politifact's fact-checks are evenly distributed across topics. Fact Checker focuses more on salient topics, such as the Clinton controversy and Healthcare.

Next, to observe if fact-checkers tend to agree more or less often in certain topic areas, I use the trained LDA to assign topic probability to 77 pairs of Overlap statements and 56 pairs of Murky statements. Each pair consists of direct quotes of political statements that were fact-checked by both Politifact and Fact Checker. Unlike the previous set of texts used to train the LDA, the new set does not include the fact-checkers' evaluation of the statements to ensure that the two statements in a pair is assigned to a single topic. I then compute the consensus rate for Overlap statements by topic (see Table 3).

Cohen's κ coefficients are computed by topic area for both samples: an Overlap-only sample and an Overlap/Murky sample. The number of statements under each topic is very small. For Immigration, Campaign, and Economy-related statements, weighted Cohen's κ coefficients exceed 0.7, a commonly applied standard in intercoder reliability tests. Because these statements usually involve statistical figures, such as polling results, the unemployment rate, or the number of refugees, there may have been relatively little room for disagreement between fact-checkers.

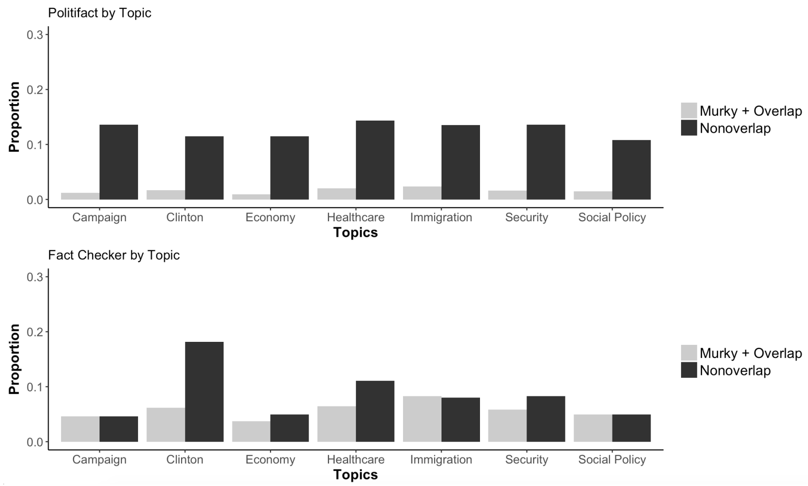


Figure 3 - Proportion of fact-checks by topic The 7 LDA-classified topics are on the x-axis and the level of attention dedicated to each topic is plotted on the y-axis.

In contrast, both unweighted and weighted κ coefficients are below or close to zero for Social Policy (e.g., candidates' views on education and job creation policies) and the Clinton controversy statements. Intercoder reliability for the Clinton controversy statements (e.g., the email controversy and accusations against the Clinton Foundation) may be particularly low due to the nature of the controversy itself. In many cases, while both fact-checkers agree that the statement is inaccurate, they disagree on how inaccurate it is, leading to discrepancies in ratings.

4 Conclusion

I examine two of the main factors which fact-checkers suggest make fact-checking powerful: 1) large overlap and 2) high consensus rate among major fact-checkers. My findings suggest that fact-checkers rarely fact-check the same statement and when they do, the rate of agreement on its factual accuracy is quite low for statements in the relatively ambiguous scoring range (i.e., "Half True" or "Mostly False"). Relative to outright falsehoods or obvious truths, these statements are likely to be carefully crafted, subtle forms of deceptive remarks. This is where fact-checking has the most to offer to the public. Yet, this is the scoring range in which they struggle the most. The surprisingly low rate at which different fact checkers agree

Table 3 Consensus Rates for Overlap and Murky Statements by Topic
Table 3a. shows consensus rates for Overlap statements only. Table 3b. shows consensus rates for both Overlap and Murky statements. Column 1 is a list of topics. Columns 2 and 3 show unweighted and weighted Cohen's κ coefficients, respectively. Column 4 shows the number of statements in each topic area.

Table 3a. Overlap Statements Only
Intercoder reliability among Immigration and Economy statements are high whereas Cohen's κ for Social Policy and Clinton statements are quite low.

Topic	Unweighted κ	Weighted κ	N
Immigration	0.656	0.951	18
Social Policy	−0.190	−0.400	5
Security	0.480	0.649	13
Campaign	0.489	0.773	12
Healthcare	0.312	0.500	11
Clinton	0.118	0.079	9
Economy	0.534	0.905	9

Table 3b. Overlap and Murky Statements
Weighted κ for Immigration and Campaign statements are high while intercoder reliability among Social Policy and Healthcare statements are very low.

Topic	Unweighted κ	Weighted κ	N
Immigration	0.464	0.869	25
Social Policy	0.023	0.243	13
Security	0.196	0.600	20
Campaign	0.509	0.818	20
Healthcare	0.159	0.422	28
Clinton	0.305	0.400	15
Economy	0.345	0.545	12

when evaluating the same statements in this scoring range suggests that providing objective information about candidates' honesty is quite difficult.

To better serve as a second pair of eyes for one another, fact-checkers should be more specific about what counts as "factcheck-worthy". A more specific definition of "factcheck-worthy" may enable multiple fact-checkers to fact-check a common set of key political claims more frequently. As a result, readers will benefit from having multiple fact-checkers independently evaluate the same statement. Also, although fact-checkers do not use directly comparable scales, fact-checkers may still improve consistency in their ratings by simply reporting the types of logical fallacies without assigning scores to these observations.

Improving consistency in how different fact-checkers choose and evaluate political claims will make fact-checking more effective and thereby help fact-checkers fulfill the democratic ideal of the political watchdog by preventing political lying.

Notes

[1] Example: https://tinyurl.com/ydg8mnsy.

[2] Politifact publishes far more per day than Fact Checker, thanks to its state affiliates.

[3] There are 3 cases in which Politifact fact-checks each part of the claim separately in two entries while Fact Checker fact-checks the entirety of the claim in a single entry. Thus, Politifact has 3 more fact-checks.

[4] When fact-checks by Politifact's state affiliates are excluded, its overlap rate increases to 0.073.

[5] Kessler writes as follows in his email correspondence with Marietta et al. (2015) on March 24, 2014: "This is how I view it: Geppetto=true, One Pinocchio=mostly true, Two Pinocchios =half true, Three Pinocchios=mostly false, Four Pinocchios=false, Pants on Fire."

References

Amazeen MA (2013) Making a difference: A critical assessment of fact-checking in 2012. *New America Foundation Media Policy Initiative Research Paper.*

Amazeen MA (2016) Checking the Fact-Checkers in 2008: Predicting Political Ad Scrutiny and Assessing Consistency. *Journal of Political Marketing* 15:4, 433-464.

Barrett P (2001) Interrater Reliability: Definitions, Formulae, and Worked Examples.
Available at: http://www.pbarrett.net/presentations/rater.pdf (accessed 7 March 2018)

Froomkin D (2012) How the mainstream press bungled the single biggest story of the 2012 campaign. *Huffington Post*
Available at: `https://www.huffingtonpost.com/dan-froomkin/republican-lies-2012-electio`
`b_2258586.html` (accessed 9 March 2018)

Gottfried JA, Hardy BW, Winneg KM and Jamieson KH (2013) Did Fact Checking Matter in the 2012 Presidential Campaign? *American Behavioral Scientist*, 57(11):1558-1567.

Graves L (2016) *Deciding What's True: The Rise of Political Fact-checking in American Journalism.* Columbia University Press.

Kessler G (2016) Trump repeats the same lines, even when they're proven false. *The Charlotte Observer*,
Available at: `http://www.charlotteobserver.com/opinion/article76441787.html`
(accessed 9 March 2018)

Marietta M, Barker DC and Bowser T (2015) Fact-Checking Polarized Politics: Does The Fact-Check Industry Provide Consistent Guidance on Disputed Realities? *The Forum* 13(4): 576-596.

Nyhan B and Reifler J (2015) The Effect of Factchecking on Elites: A Field Experiment on US State Legislators. *American Journal of Political Science* 59(3): 628-640.

Shepsle KA (1972) The Strategy of Ambiguity: Uncertainty and Electoral Competition. *American Political Science Review* 66(2):555-568.

Waldman, P (2015) Why Donald Trump Is Impervious to Fact-checking. *The Week*, December 1.
Available at: `http://theweek.com/articles/591476/why-donald-trump-impervious-factcheck`
(accessed 9 March 2018)

Can Fact-checking Prevent Politicians from Repeating Falsehoods?

Chloe Lim

August 28, 2020

Abstract

Journalists now regularly trumpet fact-checking as an important tool to hold politicians accountable for their public statements, but fact-checking's effects on politicians have only been assessed anecdotally and in experiments on politicians holding lower-level offices. Using a rigorous research design to estimate the effects of fact-checking on presidential candidates, this paper shows that a fact-checker deeming a statement false is associated with a 9.3 percentage points reduction in the probability that the candidate repeats the claim. To eliminate alternative explanations that could confound this estimate, I use two types of difference-in-differences analyses, each using true-rated claims and "checkable but unchecked" claims, a placebo test using hypothetical fact-check dates, and a topic model to condition on the topic of the candidate's statement. This paper contributes to the literature on how news media can hold politicians accountable, showing that when news organizations label a statement as inaccurate, they may affect candidate behavior.

Journalists now regularly trumpet fact-checking as an important facet of watchdog journalism that can hold politicians accountable for their public statements. In particular, in the wake of the current "fake news" crisis in a so-called "post-truth" era, advocates argue that fact-checking can help mitigate the spread of political misinformation by holding political elites accountable for what they say. According to the Duke Reporters' Lab database of global fact-checking websites, there are around 50 active fact-checking outlets in the United States.[1]

Due to its increasing popularity, fact-checking has received much scholarly attention. A number of Political Scientists and Communications scholars have examined if fact-checking can reduce the spread of false information among citizens. Nyhan (2010), Shin et al. (2017) and Shao et al. (2018) find that in general, fake news and misinformation are resilient to fact-checking and these propagate even after being debunked by fact-checkers. In contrast, Wood and Porter (2016) argue that citizens heed factual information, even when such information is not consistent with their ideological beliefs.

The effects of fact-checking on elite behavior have also been assessed anecdotally by journalists and campaign staffs. Additionally, Nyhan and Reifler (2015) conducted an experiment on politicians holding lower-level offices and found that fact-checking does play a role in preventing politicians from lying.

I build on the growing literature on fact-checking by using a rigorous research design to evaluate its real-world effects on subsequent candidate behavior in 2012 and 2016 presidential elections. As a baseline design, I use an interrupted time series design to show that a fact-checking agency deeming a statement false causes a 9.3 percentage point decrease in the probability that the statement is repeated in the future. To eliminate alternative explanations that could confound this estimate, I use a series of robustness checks and show that the core conclusion remains. I use two different types of difference-in-differences analysis, each using a dataset of "checkable but unchecked" statements and true-rated statements to ensure that the

[1] https://reporterslab.org/fact-checking/

conclusions from the interrupted time series design are not merely the result of politicians failing to repeat statements for reasons unrelated to the fact-checking itself. I also conduct a placebo test that evaluates the trend in false claims during randomly chosen, hypothetical dates when a statement could have been fact-checked but was not. Finally, I use a topic model to assess if the time at which presidential candidates switch to new agenda items coincides with the time of fact-check.

Across research designs, I find that fact-checking is associated with a decrease in the probability that politicians repeat false-rated claims after the fact-check. The effects were especially pronounced for candidates in 2016. For Hillary Clinton, the probability of repeating a false-rated claim was reduced by 14.67 percentage point after the fact-check. Similarly, for Donald Trump, the probability of repeating a statement in the future decreased by 10.83 percentage point once the statement was found to be false by a fact-checker.

Even if most of the public may never directly encounter the fact-checks themselves in their online news consumption (Guess 2018), this paper demonstrates an important channel through which fact-checking can matter in presidential campaigns. By showing that news organizations may affect candidate behavior by scrutinizing and evaluating their public statements, this study contributes to the literature on how news media can hold politicians accountable.

Mixed Evidence on the Effects of Fact-checking

A press that actively covers and scrutinizes political figures helps keep the quality of democratic governance in check (Snyder 2010). As a political watchdog, media provides voters with political information, such as how elected officials are performing in Congress or if candidates have made dishonest appeals in campaign advertisements[2]. An active media coverage and scrutiny help voters hold politicians accountable, which could in turn affect electoral outcomes and improve legislative

[2] Ad watch, the media's scrutiny of candidates who make deceptive appeals in campaign advertisements, hurt the reputation of these candidates among voters (Min 2002, Cappella and Jamieson 1994, Sullivan and Geiger 1995)

behavior (Snyder and Strömberg 2010, Ansolabehere and Iyengar 2006).

Fact-checking, a new form of media scrutiny, gained prominence as a tool to increase accountability among political figures by punishing those who distort the truth (Graves 2016). But how effective is fact-checking in holding political figures accountable for their words?

More specifically, do politicians respond to negative fact-checks by abandoning claims that are proven to be false? Anecdotal evidence on the effects of fact-checking on politicians is mixed. It appears that fact-checking did have an impact on politicians at the state and local level. In 2012, a candidate running for the Ohio State Senate who earned a lot of "False" and "Pants on Fire" from Politifact lost the race when voters turned away from him citing his dishonesty (Graves 2016). Stencel (2015) also writes about a couple of instances in which politicians modified their rhetoric after seeing a negative fact-check. To avoid being called out for dishonesty, a number of campaigns have even appointed staff members to deal with fact-checkers by lobbying to get an "advance clearance" on claims before the statements are out in the field (Stencel 2015).

In contrast, anecdotal evidence suggests that fact-checking has not been as effective at the presidential level. Mark McKinnon, a strategist for former President George W. Bush, described the presence of fact-checkers in the campaign season as "it's like everyone is driving 100 miles per hour in a 60-miles-per-hour zone and all the cops have flat tires" (Carr 2012). According to Carr (2012), a former journalist for New York Times, despite being fact-checked by multiple news organizations and ad watches, candidates for the 2012 election kept repeating false statements even after fact-checkers found these statements to be false. Even the fact-checkers themselves make only very modest claims about their impact on political figures. According to Graves (2016), while most fact-checkers can cite cases in which a politician dropped a talking point once it was rated "false", they still concede that "political lying continues unabated and always will" due to a widespread disregard for truth among politicians.

Due to a lack of empirical analyses and mixed anecdotal evidence on the real-world effects of fact-checking on elite behavior, the question of whether fact-checking can effectively monitor political figures remains unanswered. Few studies have formally assessed the impact of fact-checking on political elites. The current literature on fact-checking is largely focused on the effects of fact-checking on voter behavior. These papers examine if fact-checking can alter or reinforce voters' perception of political figures' issue stands or trustworthiness. Drawing on survey responses, Gottfried et al. (2013) found that fact-checking has been effective in improving the accuracy of voters' perception of candidates' policy platforms. Also, Wintersieck (2017) showed that candidates whose dishonesty was exposed by a fact-checking outlet received negative evaluations among survey respondents.

Then, how do politicians respond to anticipated changes in voters' perception of their honesty after seeing a negative fact-check? Nyhan and Reifler (2015) conducted a field experiment in which state legislators from nine US states received letters warning them of electoral consequences of receiving a negative rating from a fact-checker. Nyhan and Reifler (2015) found that legislators who had received the letter were less likely to make false-rated claims compared to those who had not received the letter.

Here is how my design differs from Nyhan and Reifler (2015)'s approach: Nyhan and Reifler (2015) evaluate the effects of informing state legislators of the *possibility* of being fact-checked without actually having any of their statement evaluated by a fact-checker. In contrast, the presidential candidates in my study received actual dishonesty ratings (as opposed to warnings) from a fact-checker, which were then made publicly available online. My design allows me to examine whether an *actual practice* of fact-checking affects politicians' tendency to repeat a false-rated claim in subsequent speeches.

Using an interrupted time series analysis, I observe how (or if at all) politicians modify their behavior when they are called out by fact-checkers for being inaccurate or misleading. The remainder of this paper is organized as follows. The next section describes the data collection procedure in detail. Then, I show that fact-checking is associated with a reduction in the probability that politicians repeat false-rated claims. To validate these findings, I address two alternative hypotheses according to which changes in candidates' behavior found in the main model may be caused by factors that are completely unrelated to fact-checking, such as "topic switch" or "natural decrease over time" (see Testing Alternative Explanations Section for detailed explanation). Using two types of difference-in-differences analyses, robustness checks using placebo treatment dates, and an unsupervised classification model, I find that these alternative hypotheses alone fail to explain the drop in the number of false-rated claims after the fact-check and that fact-checking may have played a role in deterring presidential candidates from repeating false claims. In addition, I show that the effectiveness of fact-checks does not depend on the saliency of campaign events in which a fact-checked statement is made. The final section offers several possible explanations for the results.

Collection of Fact-checked and Unchecked Statements

I gathered 361 speeches (Clinton: 67, Obama: 105, Romney: 77, Trump: 112) made by presidential candidates for the US Presidential elections of 2012 and 2016 between August 1st and November 5th (for 2012) / November 7th (for 2016).[3] These include speeches or remarks made in campaign rallies, presidential debates, and national conventions (See Appendix I for a complete list of speeches used in my analysis). I chose to focus on these speeches, because they are the main channel through which candidates appeal to voters by introducing their campaign promises and issue platforms. The speech transcripts were available on C-SPAN and the American Presidency Project at UC Santa Barbara.

I focus on the period between early August and Election Day to evaluate the

[3] Romney and Obama each had 79 and 116 speeches originally. However, 13 speeches were excluded because they were exclusively about exogenous events that occurred during the campaign – attack on the U.S. Consulate in Benghazi and natural disasters such as Hurricane Sandy, Tropical Storm Isaac, and unusual heat.

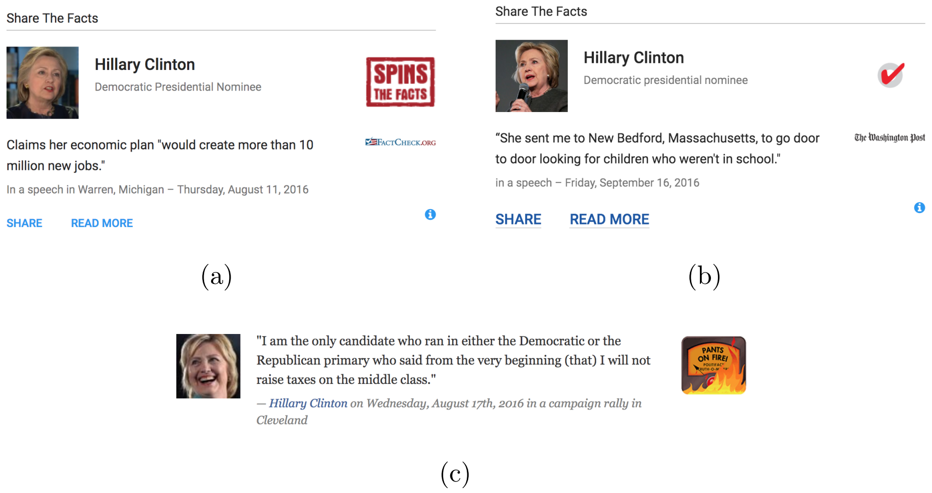
immediate effects of fact-checking more precisely than including the earlier months of the year. From January to late July of each election year, the average number of days between two consecutive speeches was 6.321. For instance, during a town hall meeting held on April 21, 2016, Clinton made a claim about her record during the 2008 primary. That afternoon, this claim was evaluated by Politifact. I then observe whether or not Clinton repeated this claim in 5 subsequent speeches after the date of fact-check, which were delivered on the following dates: April 22, April 26, May 3, May 10, and May 26. In this case, the 5 post-factcheck speeches were made over the span of 35 days and by the time Clinton delivered her May 26 speech, more than a month had passed after the date of fact-check. Thus, whether Clinton decides to repeat a fact-checked claim or not could be a result of other political events or factors that are irrelevant to fact-checking.

Yet, between August and early November, speeches are made almost daily in rallies and conventions (0.98 speech per day on average). Trump's claim about Clinton's child care plan (made during a campaign event in Aston, PA on September 13) was fact-checked on the same day by Politifact. Following the fact-check, the subsequent 5 speeches were made on September 14, September 15, September 16[4], and September 17. By the time Trump delivered his September 17 speech, it had only been 4 days since the date of fact-check. Observing the immediate effects of fact-checking on candidate behavior in post-factcheck speeches over a much shorter time span makes it less likely for other political events or actions to confound the analysis.

The dataset consists of 292 fact-checked statements and 142 unchecked statements. Each fact-checked statement corresponds to a direct quote of the statement that has been fact-checked by a fact-checker. Figure 1 shows examples of a direct quote of a fact-checked statement from each of the three online major fact-checking outlets. Likewise, each unchecked statement represents a direct quote of the candidate's statement that has not been fact-checked by a fact-checker. Unchecked

[4] Two speeches were made on September 16, 2016.

Figure 1. Examples of a "Direct Quote" of a Fact-checked Statement
For the majority of fact-checks, a direct quote of the statement that is being fact-checked is presented in a box, along with the source (i.e., name of the corresponding fact-checking outlet) and a rating. (a), (b), (c) each represent a fact-checked statement by FactCheck.org, Washington Post's Fact Checker, and Politifact, respectively. Whenever a fact-checked statement is not presented in this format, the statement is either displayed as bullet-point list or is introduced at the beginning of the article.



Share The Facts

Hillary Clinton
Democratic Presidential Nominee

Claims her economic plan "would create more than 10 million new jobs."

In a speech in Warren, Michigan – Thursday, August 11, 2016

(a)

Share The Facts

Hillary Clinton
Democratic presidential nominee

"She sent me to New Bedford, Massachusetts, to go door to door looking for children who weren't in school."

in a speech – Friday, September 16, 2016

(b)

"I am the only candidate who ran in either the Democratic or the Republican primary who said from the very beginning (that) I will not raise taxes on the middle class."

— *Hillary Clinton on Wednesday, August 17th, 2016 in a campaign rally in Cleveland*

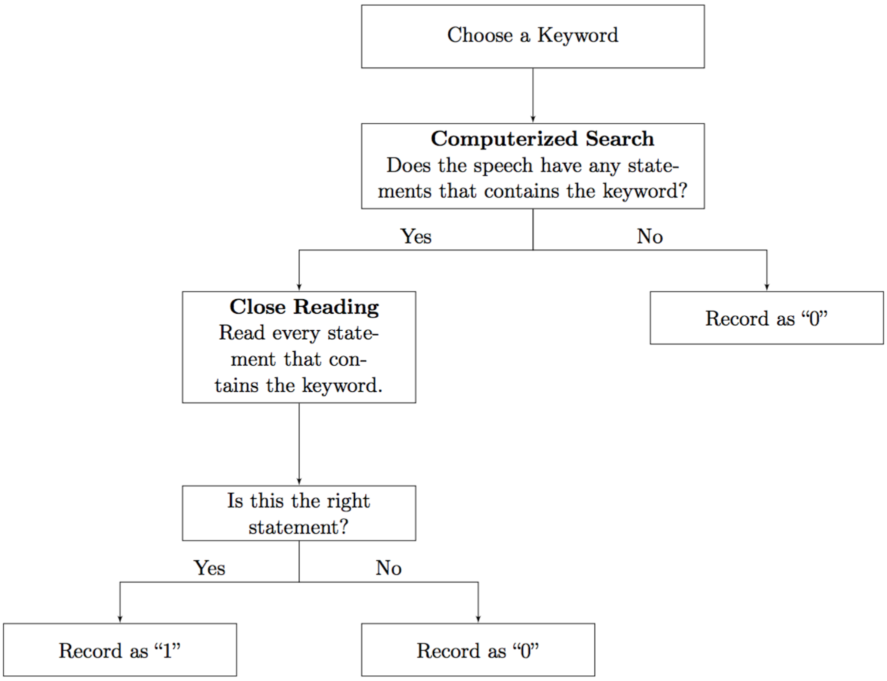
(c)

statements were collected from the 361 speeches in my sample.

The unit of analysis is a statement - speech pair. Each unit takes the value of 1 if a particular statement was made in a given speech, and 0 otherwise. I also collected ClaimBuster scores for every statement. ClaimBuster scores indicate "checkability", which is determined by whether or not a given sentence contains verifiable, factual claims.[5] The score ranges from 0.0 to 1.0. The higher the score, the more factual and "factcheck-worthy" the sentence is. (See Appendix B for a list of fact-checked statements and their ClaimBuster scores.)

The next two sections describe the data-collection process for fact-checked and unchecked statements, respectively.

Figure 2. A Two-Step Procedure for Finding Fact-checked Statements


Choose a Keyword
Computerized Search
Does the speech have any statements that contains the keyword?
Yes
No
Close Reading
Read every statement that contains the keyword.
Record as "0"
Is this the right statement?
Yes
No
Record as "1"
Record as "0"

Collecting and Recording the Frequency of Fact-checked Statements

First, I gathered "fact-check articles" from three online major fact-checkers – FactCheck.org, Politifact, and Washington Post's Fact Checker during the 2012 and 2016 general election campaigns.[6] Each "fact-check article" consists of a direct quote of the statement being fact-checked and the fact-checker's evaluation of the statement. While Politifact and Washington Post's Fact Checker evaluate a single statement per "fact-check article", FactCheck.org often evaluates multiple statements at once in a single article. From each of these "fact-check articles", I collected a direct quote of the statement that is being fact-checked and created a dataset of 292 fact-checked statements.

The next task was to record whether or not a given fact-checked statement was

[5] ClaimBuster is a fact-checking platform created by Hassan et al. (2017). The ClaimBuster scores are obtained from a supervised classification model. The model was trained using 20,788 sentences from past general election debates, which were labeled by human coders (Hassan et al. 2017).

[6] Washington Post's Fact Checker does not have fact-checks for the 2012 presidential election. Therefore, only FactCheck.org and Politifact are used for Obama and Romney.

made in each speech. I carried out the search in two stages – a preliminary computerized search and a careful, close reading for those that passed the preliminary round (Figure 2 illustrates the two-step procedure for finding fact-checked statements). To avoid false negatives (i.e., failing to note that a fact-checked statement is included in a given speech), I intentionally chose a very vague, stemmed keyword for each search. For example, the keyword used for statements on immigrants or immigration policy was "immig". Also, whenever there was a common synonym for the keyword (e.g., "job loss" and "unemployment"), I ran multiple rounds of search with different search terms.

Then, every speech that contained a given keyword moved on to the next stage. Because the preliminary computerized search returned a lot of false positives due to intentionally vague, stemmed keywords, I read all speeches that made it to the second round and recorded whether a given statement was made in each speech. Through a close reading of speeches, I eliminated instances in which a keyword appears in a speech, but in a different statement or context.

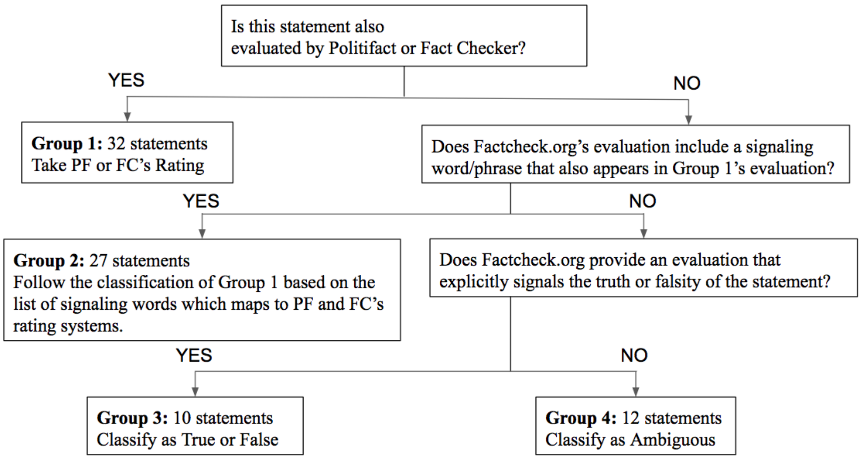
Classification of Ratings

Both Politifact and Fact Checker assign ratings for claims based on their own rating scales. Politifact has a 6-point scale: "True", "Mostly True", "Half True", "Mostly False", "False", and "Pants on Fire". Fact Checker uses a 5-point scale: "Geppetto Checkmark", "1 Pinocchio", "2 Pinocchios", "3 Pinocchios", and "4 Pinocchios". Following Kessler[7]'s interpretation of how Fact Checker's scale compares to that of Politifact, I group Geppetto Checkmark with True, 1 Pinocchio with Mostly True, 2 Pinocchios with Half True, and 3 Pinocchios with Mostly False, and both False and Pants on Fire with 4 Pinocchios.

For my analysis, I re-classify the ratings into 3 categories: "True", "Ambiguous", and "False". In my classification, "True" includes "True" and "Mostly True" from

[7] Glenn Kessler, the Chief Editor at Fact Checker, writes as follows in his email correspondence with Marietta et al. (2015) on March 24, 2014: "This is how I view it: Geppetto=true, One Pinocchio=mostly true, Two Pinocchios =half true, Three Pinocchios=mostly false, Four Pinocchios=false, Pants on Fire."

Figure 3. Assigning Scores to Factcheck.org's Fact-checks


Is this statement also evaluated by Politifact or Fact Checker?
YES
NO
Group 1: 32 statements
Take PF or FC's Rating
Does Factcheck.org's evaluation include a signaling word/phrase that also appears in Group 1's evaluation?
YES
NO
Group 2: 27 statements
Follow the classification of Group 1 based on the list of signaling words which maps to PF and FC's rating systems.
Does Factcheck.org provide an evaluation that explicitly signals the truth or falsity of the statement?
YES
NO
Group 3: 10 statements
Classify as True or False
Group 4: 12 statements
Classify as Ambiguous

Politifact and "Geppetto Checkmark" and "1 Pinocchio" from Fact Checker. "False" includes "Mostly False", "False", and "Pants on Fire" from Politifact and "3 Pinocchios" and "4 Pinocchios" from Fact Checker. "Ambiguous" includes Politifact's "Half True" and Fact Checker's "2 Pinocchios".

Unlike Politifact and Fact Checker, Factcheck.org does not assign numerical scores to fact-checked statements. Instead, in many cases, it concludes with a phrase or a word that implies whether a given statement is closer to being true or false. Sometimes, its conclusion can be as explicit as labeling the statement "correct" or "false". Figure 3 describes how I converted and classified Factcheck.org's evaluations into 3 categories.

First, as shown in Figure 3, I check if the statement evaluated by Factcheck.org was also fact-checked by Politifact or Fact Checker. Fortunately, about 40 percent of the statements (32 out of 81) that are evaluated by Factcheck.org are also fact-checked by either Politifact or Fact Checker (or both). These statements were classified based on their rating on either Politifact or Fact Checker. For instance, if Statement A is fact-checked by both Factcheck.org and Politifact, it takes Politifact's rating.

For claims that were evaluated by both Factcheck.org and Politifact/Fact Checker, I compared Factcheck.org's evaluation with that of Politifact or Fact Checker and found that, in general, if Factcheck.org's evaluation of a statement included any of the following phrases, it received "False" on either Politifact or Fact Checker: "falsely claimed", "cherry-picks", "misleading", "wrongly said", "false", "distorts the facts", "questionable", "bogus", "no evidence", "outdated figure". Statements that received "Half True" / "2 Pinocchios" on either Politifact or Fact Checker were usually evaluated as follows on Factcheck.org: "missing context", "exaggerated", "goes too far", or "straining facts". Factcheck.org's conclusion for claims that received "True" on Politifact and Fact Checker's scale included sentences such as "[Candidate Name] was right/correct". (See Appendix G for a complete list of statements with ratings and explanations from Politifact/Fact Checker and Factcheck.org.)

The rest (49 statements) were fact-checked by Factcheck.org alone. I assigned scores to 27 of them, using the aforementioned list of key phrases which implied what the corresponding rating on Politifact or Fact Checker would be. Of the remaining 22 statements, there were 10 cases in which Factcheck.org explicitly called a given statement incorrect/unsupported or provided reasons for why the statement is wrong or biased. Therefore, these 10 statements were labeled "False". The remaining 12 cases were a bit less straightforward. For example, the evaluation would call the statement "not the whole story", "technically correct but misleading" or "true but the way he framed it was a stretch". Because I wanted to avoid misclassifying statements with ambiguous ratings as "False", I classified these 12 as "Ambiguous". Interrater Reliability Rate among two independent coders on 50 randomly selected fact-checked statements on FactCheck.org was 0.75, which is well within the 0.7-0.8 range, the standard of interrater reliability that academics commonly apply when evaluating hand-coded data (Barrett 2001).

Dataset of Unchecked Statements

I collected "checkable but unchecked" statements according to a rule fact-checkers have established for themselves about what counts as "checkable" : "facts, not opinions". In collecting "checkable but unchecked" statements, I relied on Graves' explanation of examples and counterexamples of “checkable” statements provided by editors at training sessions for interns at FactCheck.org and reporters for Politifact’s state franchises.

Then, I closely read all speeches and first picked out every unchecked statement, with the exception of normative statements (e.g., “You should be able to serve your country no matter who you are.”), unverifiable predictions (e.g., “Mexico will pay for the wall.”), opinions, or statements that contained wishes or emotions (e.g., “I am certainly relieved that my father never did business with Donald Trump.”). Later, in estimating difference-in-differences estimates comparing unchecked vs. checked claims, unchecked statements are paired with fact-checked statements based on similarity in topic, pre-treatment (fact-check) frequency, and ClaimBuster scores. Thus, even if a few relatively less “checkable” statements were included in the dataset of unchecked statements, these statements will have received ClaimBuster scores that are too low to be matched with any of the fact-checked statements and hence will be dropped from the analysis.

In addition to being “checkable”, fact-checkers note that they look for statements about an important and salient policy matter or claims that highlight differences among candidates (“usually accusations leveled by one candidate against another”) (Graves 2016). These rules were relatively easy to follow, because this paper focuses on speeches made by presidential candidates 1-3 months prior to the election. Thus, with the exception of a handful of “checkable” but policy-irrelevant statements, such as Trump’s remarks about an attire worn by a man who attended one of his speeches, most “checkable” statements were claims that have received media attention, claims about important policy matters, or accusations aimed at another candidate.

Next, I applied a two-stage keyword search (the same procedure used for fact-checked statements (see Figure 2)) for the collection of “checkable but unchecked” statements. Statements that were fact-checked in the past, but not during the period of interest, were taken out of the sample (See Appendix C for a complete list of unchecked statements).

Evaluating the Effect of Fact-checking with Interrupted Time Series Design

To analyze the effect of fact-checking on the likelihood that presidential candidates will repeat a fact-checked claim, I use an interrupted time series design to compare the frequency with which each statement was made before and after being fact-checked. The date of fact-check divides the speeches into a treatment and a control group. Speeches that were given before the day of fact-check are the control group and speeches that were made after the fact-check are the treatment group.

Speeches that were made on the day of fact-check are excluded, because in most cases, it is unclear whether fact-checks are published before or after the candidate's speech of the day. Depending on the venue or the type of event, the time of day at which candidates make remarks highly varies. Speeches can be given in the morning, early afternoon, late afternoon, evening (e.g., fundraising dinners), or at night (e.g., presidential debates). Moreover, of the three major fact-checking outlets used in my analysis, only Politifact has timestamps on its articles. Fact Checker and FactCheck.org only display dates. Typically, on Politifact, fact-checks are published at various times throughout the day from early in the morning to late at night. Thus, it is impossible to designate a specific time at which fact-checks usually occur and interpolate this information to Fact Checker and FactCheck.org.

For each fact-check, the control group includes 5 speeches that immediately precede the day of fact-check (for convenience, pre-factcheck speeches are labeled "Speech -5", "Speech -4", "Speech -3", "Speech -2", "Speech -1") and the treatment group has 5 speeches that immediately follow the day of fact-check (for convenience, post-factcheck speeches are labeled "Speech 1", "Speech 2", "Speech 3", "Speech 4", "Speech 5"). Because different statements are fact-checked on different dates and speeches are chosen relative to the date of fact-check, each statement has its own set of "Before" and "After" speeches. For example, "Speech 1" for Statement A and "Speech 1" for Statement B may be two different speeches made on different dates.

However, for convenience, I use the normalized label "Speech n" to refer to a collection of "Speech n" for all statements (See Appendix A for an example of what the dataset looks like).

On average, a set of 5 speeches is given over the span of 5.4 days, which will allow for sufficient time for candidates and their campaign staff to learn about the fact-check and if necessary, make changes to their subsequent speeches accordingly. Also, restricting the treatment group to 5 speeches after the fact-check enables me to compute the immediate effects of fact-checking over a relatively short time span and therefore reduces the possibility that the candidate's choice of whether or not to repeat a fact-checked claim is confounded by other political events or actions. Expanding the size of the treatment group would require me to observe changes in candidate behavior over a longer time span, which may confound the estimate. Hence, I focus on 5 post-factcheck speeches.

To estimate the effect of fact-checking, I use a fixed effects regression given by

$$Spoken_{it} = \eta_i + \alpha Factcheck_{it} + \beta Length_{it} + s_j(o_t) + \epsilon_{it} \tag{1}$$

Spoken is coded as 1 if Statement i appears in Speech t, and 0 otherwise. To ensure that I rely on variation within each statement, I include η_i, a statement fixed effect which rules out omitted variable bias from unobserved statement-specific characteristics that do not vary across speeches. $Factcheck_{it}$ is coded as 1 if a given Speech t is made after the day in which Statement i is fact-checked, and 0 otherwise. For example, for a given statement i, *Factcheck* for Speech t is 1 for $t = 1, 2, 3, 4, 5$ and 0 for $t = -5, -4, -3, -2, -1$. The quantity of interest is α, which represents the change in the probability that a particular statement appears in speeches once it is debunked by a fact-checker. I control for $Length_{it}$, the length of each speech (measured by the number of words in each speech), which may affect candidates' decision to make a particular statement or not, independent of fact-check (i.e., in general, candidates will say more during a longer speech and therefore, the probability that any given fact-checked claim will appear in longer speeches is higher, compared to shorter speeches). Standard errors are clustered at the statement level. $s_j(o_t)$ are various functions that model time trends using the running variable o_t – the order of speech relative to the time of fact-check. For example, $o_t = 2$ for "the second speech after the time of fact-check" and $o_t = -4$ for "the fourth speech before the time of fact-check". Following the specification used in Mummolo (2018), the model estimates either a simple difference in means (in which case $s_j(o_t)$ is simply omitted from the equation) or linear, quadratic, or cubic functions by interacting various orders of o_t with $Factcheck_{it}$.[8]

Do Candidates Avoid Repeating False-Rated Claims?

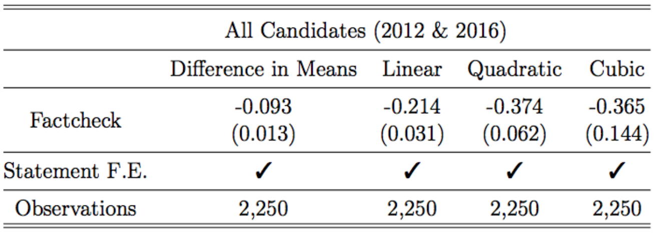
Table 1 shows the interrupted time series estimate for the effects of negative ("false") ratings from a fact-checker on the probability that candidates from the 2012 and 2016 presidential elections make false-rated claims. On average, fact-checking is associated with a 9.3 percentage point decrease in the probability that a candidate will repeat a false-rated claim. Perhaps because fact-checking became more popular and frequent in 2016 compared to the previous election[9], the effect of fact-checking is more pronounced for candidates who ran in 2016, relative to those in 2012. Receiving a "False" from fact-checkers is associated with a 9.8 percentage point decrease in the probability that candidates in 2016 repeated a given statement. For candidates in 2012, although the coefficient on *Factcheck* is negative, the effect size is smaller, compared to their 2016 counterparts. As shown in Table 1, the estimates are negative and significant across a variety of specifications.

Figure 4 displays the percentage of false-rated statements made in each speech, relative to the day of fact-check. The percentage of false-rated claims per speech is computed as follows: $\frac{\sum_{i=1}^{n} Spoken_i}{n}$ where n represents the total number of false-

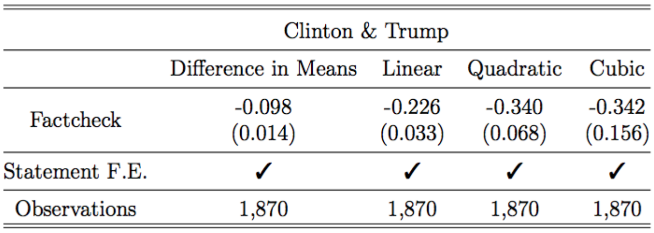
[8] The linear model is specified as $Spoken_{it} = \eta_i + \alpha Factcheck_{it} + \beta Length_{it} + \gamma_1 o_t + \gamma_2 o_t Factcheck_{it} + \epsilon_{it}$. The quadratic model is specified as $Spoken_{it} = \eta_i + \alpha Factcheck_{it} + \beta Length_{it} + \gamma_1 o_t + \gamma_2 o_t Factcheck_{it} \gamma_3 o_t^2 + \gamma_4 o_t^2 Factcheck_{it} + \epsilon_{it}$. The cubic model is specified as $Spoken_{it} = \eta_i + \alpha Factcheck_{it} + \beta Length_{it} + \gamma_1 o_t + \gamma_2 o_t Factcheck_{it} \gamma_3 o_t^2 + \gamma_4 o_t^2 Factcheck_{it} + \gamma_5 o_t^3 + \gamma_6 o_t^3 Factcheck_{it} + \epsilon_{it}$.

[9] Between August and early November of the election year, there were approximately 90 more fact-checks during 2016 than in 2012.

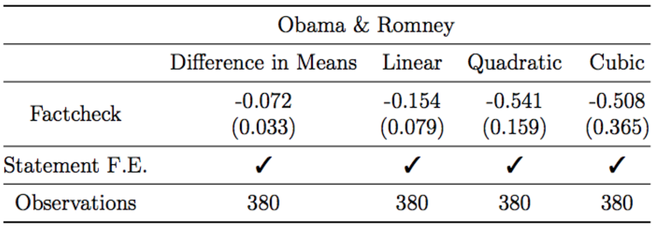
Table 1. Interrupted Time Series Estimates for False Claims with Various Specifications using Time Trends Table 1 displays the interrupted time series estimate for the effects of fact-checking on the probability that a candidate will make false-rated claims. Column 1 of each table represents estimates from a simple difference in means equation (in which case $s_j(o_t)$ is omitted from the equation). Columns 2, 3, and 4 each displays estimates from equations using linear, quadratic, and cubic functions of the time trend (o_t) and interactions of various orders of o_t with $Factcheck_{it}$. On average, fact-checking is associated with a 9.3 percentage point decrease in the probability that a candidate will repeat a false-rated claim. The estimates are negative and significant across a variety of specifications.



All Candidates (2012 & 2016)				
	Difference in Means	Linear	Quadratic	Cubic
Factcheck	-0.093 (0.013)	-0.214 (0.031)	-0.374 (0.062)	-0.365 (0.144)
Statement F.E.	✓	✓	✓	✓
Observations	2,250	2,250	2,250	2,250



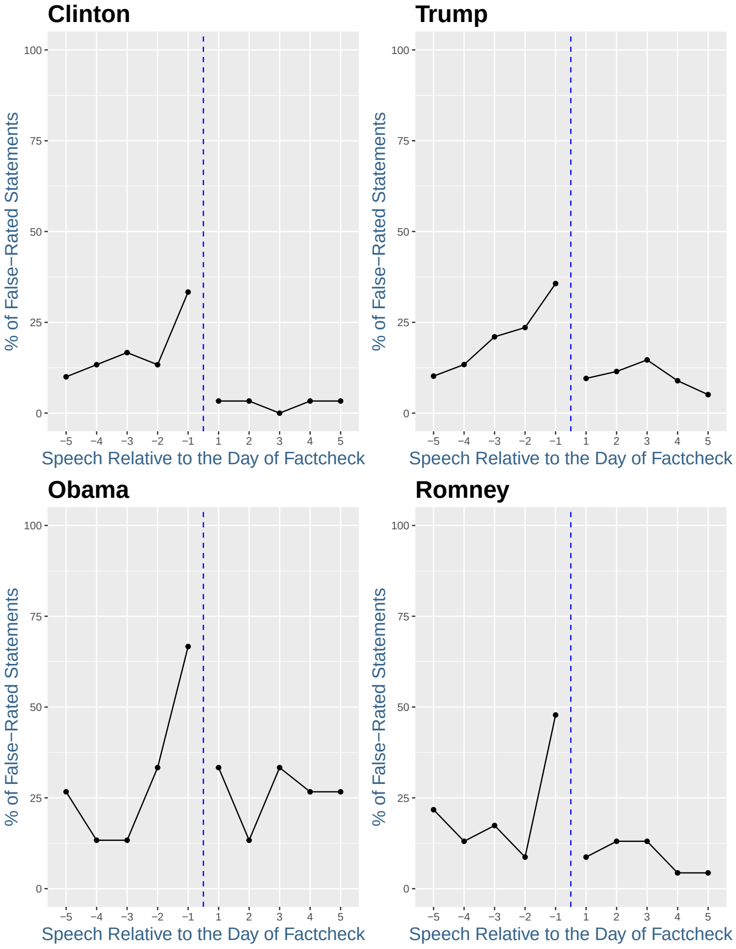
Clinton & Trump				
	Difference in Means	Linear	Quadratic	Cubic
Factcheck	-0.098 (0.014)	-0.226 (0.033)	-0.340 (0.068)	-0.342 (0.156)
Statement F.E.	✓	✓	✓	✓
Observations	1,870	1,870	1,870	1,870



Obama & Romney				
	Difference in Means	Linear	Quadratic	Cubic
Factcheck	-0.072 (0.033)	-0.154 (0.079)	-0.541 (0.159)	-0.508 (0.365)
Statement F.E.	✓	✓	✓	✓
Observations	380	380	380	380

rated statements for a given candidate and $Spoken_i = 1$ if statement i appears in a given speech and 0 otherwise. The vertical dashed line in the middle indicates the day of fact-check. On the x-axis, speeches that were made before the day of fact-check ("Speech -5" through "Speech -1") are placed on the left side of the vertical line and speeches that were made after the day of fact-check ("Speech 1" through "Speech 5") are on the right side of the vertical line. As shown in Figure 4, "Speech -1" has the highest value of $\frac{\sum_{i=1}^{n} Spoken_i}{n}$ for all candidates. This implies that speeches made immediately before the day of fact-check contained the highest percentage of

Figure 4. Percentage of False Statements Per Speech Pre- vs. Post-Factcheck Figure 4 displays the percentage of false-rated statements made in each speech, relative to the day of fact-check. On the x-axis, speeches that were made before the day of fact-check ("Speech -5" through "Speech -1") are placed on the left side of the vertical line and speeches that were made after the day of fact-check ("Speech 1" through "Speech 5") are on the right side of the vertical line. The percentage of false-rated claims per speech dropped once fact-checkers proved them to be false. On average, after the fact-check, the proportion of false-rated claims in each speech (relative to the day of fact-check) decreased by approximately 12.75 percentage points in 2016 and by 8.5 percentage points in 2012.



false-rated claims. In contrast, I observe a downward spike in the percentage of false-rated statements immediately after these claims are proven "false" by fact-checkers. This implies that candidates tended to avoid repeating debunked claims in speeches that were made immediately after the day of fact-check.

In particular, during the 2016 election, on average, the proportion of false-rated claims per speech (relative to the day of fact-check) dropped by 12.75 percentage points after the fact-check. More specifically, for Clinton, an average percentage of false-rated claims per speech went down from approximately 17 to 2 once these statements were found to be false. For Trump, the average percentage of false-rated claims per speech dropped from roughly 21 to 10 after the fact-check. In 2012, for Romney, the average percentage of false-rated claims per speech decreased from approximately 22 to 9 once they were caught by fact-checkers. However, for Obama, with the exception of the steep jump which occurs immediately after the fact-check, the average percentage of false-rated claims per speech is actually higher after the fact-check (around 31), compared to before (around 27). Although the effect size varies by year, fact-checking is followed by a non-negligible drop in the percentage of speeches that feature a false claim, especially during the 2016 presidential election.

Testing Alternative Explanations

In this section, I run additional analyses to eliminate alternative explanations that could confound the results in the previous section. Each of the analyses shows that the core conclusion remains: a negative rating from a fact-checker is in fact associated with a decline in the probability that a candidate will repeat a debunked claim in subsequent speeches. Here are two potential alternative hypotheses for a decrease in the number of false statements after the fact-check.

- "Topic Switch" Hypothesis: Candidates move on to a new agenda after a certain point, which may coincide with the time of fact-check.
- "Natural Decrease over Time" Hypothesis: Candidates gradually stop making certain claims after a certain point, which may coincide with the time of fact-check.

In both cases, the candidate's decision not to repeat the statement may have nothing to do with fact-checking. If the negative coefficient on *Factcheck* were primarily an artifact of a "Topic Switch" and/or a "Natural Decrease over Time" as the above hypotheses suggest, I expect to observe the following:

- A downward spike in the number of all statements after a certain time regardless of their fact-check ratings, such as statements that are rated "True" and statements that are not fact-checked.
- A similar downward spike after any randomly chosen cutoff date among false statements.
- A significant change in content of speeches after each presidential debate, which is when 41.5 percent of fact-checks occur.

Using two different types of difference-in-differences design, placebo checks, and a topic model, the next 4 sections test each of these possibilities.

Difference-in-Differences Analysis: Fact-checked vs. Unchecked Statements

The "Natural Decrease over time" hypothesis posits that candidates gradually stop making certain claims after a certain point, regardless of whether they were fact-checked or not. According to this hypothesis, the percentage of any given "checkable but unchecked" statement should decrease after a certain point, similar to a trend observed among false-rated statements. Using a difference-in-differences design, I evaluate whether the difference in pre- vs. post-factcheck trends among fact-checked statements and unchecked statements are statistically significant.

First, each fact-checked statement was matched with unchecked statements. The three criteria used for matching were: ClaimBuster scores, pre-treatment frequency, and topics.

1. ClaimBuster Scores: As mentioned earlier, ClaimBuster scores indicate "checkability" — a criterion used by fact-checkers when looking for statements to fact-check (Graves 2016). To minimize the difference in ClaimBuster scores between a checked and an unchecked statement, nearest neighbor matching is performed.

2. Pre-Treatment (Fact-check) Frequency: Pre-treatment frequency indicates the number of times a given statement is made in a set of 5 speeches that precede the date of fact-check. Nearest neighbor matching was performed to minimize the difference in pre-treatment frequency.[10]

3. Topics: I applied latent Dirichlet allocation (LDA) to model the topics in the texts. I assume that the collection of fact-checks are driven by 4 topics, a number chosen after assessing the substantive fit within and among the clusters. Because the fact-checked statements themselves are fairly short (around 10 words on average), full articles on fact-checked statements (which consist of fact-checkers' evaluation and background information on each statement) were used as a corpus for training LDA. Then, I used the trained LDA to assign topic probability to a collection of the direct quotes of checked and unchecked statements. I also performed a manual classification by assigning each statement to one of four LDA-defined categories. For example, LDA classified Trump's statements into 4 different topics: A - Clinton Controversy and Crime; B - Immigration and Foreign Affairs; C - Economy and Campaign; D - Healthcare. Each statement was classified twice (LDA classification and manual classification based on LDA categories): $Topic$ = (LDA classification, Manual classification). A manual classification was especially helpful when LDA performed poorly. For example, according to the LDA classification, the statement "Jonathan Gruber (architect of Obamacare) said the American people are essentially stupid for approving and allowing Obamacare to happen" was assigned to Topic B - Immigration and Foreign Affairs. Under the manual classification, the statement was categorized under Topic D - Healthcare. Both classifications were used in matching.

First, an unchecked statement was matched with a checked statement with an identical set of topic pairing. In this case, $Topic_{checked} = Topic_{unchecked}$. For example, "58 percent of African American youth are unemployed." (a checked statement) and "43 percent of African-American school-aged children live in poverty" (an unchecked statement) were assigned to "Topic C - Economy and Campaign" under both LDA

[10] Cases for which one statement was not spoken at all (i.e., the number of pre-treatment speech containing the given statement: 0) during a given timeframe and the other statement is spoken once (i.e., the number of pre-treatment speech containing the given statement: 1) were excluded to avoid comparing statements that were spoken during completely different times during the campaign.

and manual classifications. Thus, these two statements were matched.

Then, nearest neighbor matching was implemented for statements with similar (but not identical) sets of topics. In this case, a topic pairing for a checked statement $Topic_{checked}$ = (LDA classification, Manual classification) and a topic pairing for an unchecked statement $Topic_{unchecked}$ = (LDA classification, Manual classification) had at least one overlapping topic. For instance, $Topic$ for "Clinton was proposing to print instant work permits for millions of illegal immigrants" (a checked statement) was (C,B). This statement was matched with an unchecked statement "Obama has allowed Syrian refugees to pour into our country" whose topic pairing was (B,B).

Based on the three criteria (ClaimBuster scores, pre-treatment frequency, and topics), matching was performed in the following order:

1. A fact-checked statement with a ClaimBuster score of x was matched with unchecked statements with ClaimBuster scores ranging from $x - 0.05$ to $x + 0.05$.

2. Compute the difference in pre-treatment frequency between the checked statement and each of the unchecked statements selected above. Discard unchecked statements whose pre-treatment frequency differs from that of the checked statement by more than 1.

3. Nearest neighbor matching was performed based on topics.

4. Fact-checked statements that failed to get matched with the closest unchecked statement after Steps 1-3, another round of nearest neighbor matching was performed with an extended ClaimBuster score range and pre-treatment frequency difference window.

In total, there are 201 matches of checked and unchecked statements. On average, a checked statement is matched with 1.2 unchecked statements. Of the 201 matches, there were 173 cases in which there was no difference in pre-treatment frequency; 24 matches differed by 1 and 4 matches differed by 2. The mean difference in ClaimBuster scores between checked and unchecked statements is 0.003

(p-value: 0.426). There were 21 cases in which the matched statements had different topic pairings but were still close in terms of ClaimBuster scores and pre-treatment frequency. Only two fact-checked statements were excluded from the analysis because I failed to find an unchecked statement within a reasonable ClaimBuster score range, pre-treatment frequency, and similar topic pairings. See Appendix J for topic classifications of checked and unchecked statements.

Next, I assign a hypothetical date of fact-check to each unchecked statement. For each unchecked statement, the hypothetical date of fact-check is the actual date of fact-check for the fact-checked statement it is matched with. For instance, if the actual date of fact-check for the fact-checked statement is September 19th, the hypothetical date of fact-check is also set as September 19th for the unchecked statement it is matched with. The hypothetical date divides the speeches into a treated and control group. Speeches that were made before the hypothetical date of fact-check are assigned to a control group and speeches that were made after the hypothetical date are assigned to a treated group.

For the difference-in-differences design, I estimate the following equation:

$$Spoken_{it} = \eta_i + \alpha Factcheck_{it} + \beta Length_{it} + \lambda_t + \epsilon_{it} \tag{2}$$

This equation is similar to (1) from the previous section, except that it includes λ_t instead of $s_j(o_t)$. λ_t is a "Time of Speech (relative to the date of fact-check)" fixed effect ($t = 1, \ldots, 10$) that is used in difference-in-differences design for unchecked versus checked statements and true versus false statements. In my analysis, speeches are selected relative to the date of fact-check. Speech n is a speech that is n speeches apart (roughly n days away) from the speech made on the date of fact-check. Time of Speech fixed effects are included to control for things that happen over time from 5 speeches preceding the date of fact-check to 5 speeches given after the fact-check.

Table 2 reports the difference-in-differences estimates for unchecked vs. checked claims with Statement and Time of Speech (relative to the date of fact-check) fixed effects for all 4 candidates and candidates for the 2016 and 2012 election, respectively.

Table 2 Difference-in-Differences Estimate for Unchecked vs. Checked Claims Table 2 displays the difference-in-differences estimates for unchecked vs. checked claims. The difference-in-differences estimates are negative, which suggests that a negative score from fact-checkers is associated with a decrease in the probability that a candidate repeats a false-rated claim.

	Unchecked vs. Checked		
	All	2016	2012
Factcheck	-0.087 (0.020)	-0.079 (0.021)	-0.120 (0.052)
Statement F.E.	✓	✓	✓
Time of Speech F.E.	✓	✓	✓
Observations	3,820	3,090	730

Speech length is included as a control and standard errors are clustered at the statement level.

The difference-in-differences estimates are negative for both 2012 and 2016, which implies that fact-checking is associated with a decrease in the probability that a candidate repeats a false-rated claim. The difference in pre- vs. post-treatment trends among fact-checked statements and unchecked statements are especially significant for candidates in 2016.

Difference-in-Differences Analysis: True vs. False Statements

According to the "Topic Switch" and "Natural decrease over time" hypotheses, the negative coefficient on *Factcheck* for false-rated claims suggests that politicians may avoid repeating false-rated statements, not because these statements are found to be false, but because they simply move onto a new agenda or no longer feel the need to re-emphasize certain claims after having already repeated them a few times in the past. Then, it would follow that regardless of ratings, the percentage of both false-rated and true-rated statements should decrease roughly at a similar rate after the fact-check.

To evaluate these hypotheses, first, an interrupted time series regression is applied to true-rated claims (See Table 3). I then compare how post-factcheck speeches

Table 3. Interrupted Time Series Estimates for True Claims with Various Specifications using Time Trends Table 3 displays the interrupted time series estimate for the effects of fact-checking on true-rated claims. Estimates from a simple difference in means equation and equations using linear, quadratic, or cubic functions of the time trend (o_t) and interactions of various orders of o_t with $Factcheck_{it}$ are presented in each column. Compared to Table 1, the magnitude of the effects are not as large across different specifications once the time trend variable is included.

	All Candidates (2012 & 2016)			
	Difference in Means	Linear	Quadratic	Cubic
Factcheck	-0.060 (0.026)	-0.065 (0.063)	-0.139 (0.127)	-0.259 (0.292)
Statement F.E.	✓	✓	✓	✓
Observations	660	660	660	660

	Clinton & Trump			
	Difference in Means	Linear	Quadratic	Cubic
Factcheck	-0.073 (0.033)	-0.046 (0.078)	0.011 (0.159)	-0.050 (0.365)
Statement F.E.	✓	✓	✓	✓
Observations	360	360	360	360

	Obama & Romney			
	Difference in Means	Linear	Quadratic	Cubic
Factcheck	-0.048 (0.043)	-0.067 (0.100)	-0.321 (0.199)	-0.488 (0.459)
Statement F.E.	✓	✓	✓	✓
Observations	300	300	300	300

differ from pre-factcheck speeches for "True" versus "False" statements.[11] The p-value for the difference in pre-treatment frequency between "True" and "False" claims is 0.382. This is evidence in favor of the parallel trends assumption.

As shown in Table 3, although the magnitude on the coefficient for true-rated claims is smaller than that for false-rated claims, the sign is negative and the effect size is pretty large. This suggests that on average, the probability of repeating a "True" statement decreased after a fact-check, even with a positive rating from fact-checkers.

It may be that a "Geppetto checkmark" from Fact Checker and "True" from

[11] "Ambiguous" statements are not used for this analysis, because it is unclear ex-ante if or how politicians would react to receiving a rating that is not so positive yet not completely false either.

Table 4 Difference-in-Differences Estimate for True vs. False Claims Table 4 displays the difference-in-differences estimates for true-rated vs. false-rated claims. The difference-in-differences estimates are small but negative, which suggests that although the effects may be small, a negative score from fact-checkers still played a role in preventing candidates from repeating false-rated claims.

	True vs. False		
	All	2016	2012
Factcheck	-0.029 (0.027)	-0.008 (0.034)	-0.053 (0.052)
Statement F.E.	✓	✓	✓
Time of Speech F.E.	✓	✓	✓
Observations	2,610	1,930	680

Politifact may have their own effects. I speculate that candidates may drop a true-rated claim even after it is found to be true, perhaps because they decide that the particular claim has successfully "gotten out there" since it has been given sufficient attention by fact-checkers. Also, because fact-checkers rarely reward candidates for heeding their approval, candidates may not feel the need to keep repeating claims that have been rated "True".

Table 4 shows difference-in-differences estimates for true-rated vs. false-rated claims with statement and Time of Speech fixed effects for all 4 candidates and candidates for the 2016 and 2012 election, respectively. Again, speech length is included as a control and standard errors are clustered at the statement level. The difference-in-differences estimates are quite small and noisy, perhaps due to an extremely small sample size relative to that of false statements.[12] Yet, the sign of the difference-in-differences coefficients are negative for all candidates. This implies that a negative score from fact-checkers may have played a role in preventing candidates from repeating false-rated claims, at least in some cases.

[12] According to Graves (2016), the three major fact-checkers focus on statements that are likely to be false. They point out that there is little point in verifying obvious truths.

Robustness Checks using Placebo Fact-check Dates

Using a placebo test, this section evaluates the “Natural decrease over time” hypothesis. Previous sections have used the actual date of fact-check to divide the speeches into a treatment and control group. In this section, I picked 10 dates (5 of which precede the actual date of fact-check and the other 5 follow the date of fact-check) as “placebo” dates of fact-check.

For each round of analysis, a corresponding “placebo” date is used to assign speeches into a treatment and control group. Because there are 10 such “placebo” dates for each fact-check, the following fixed effects regression is run 10 times:

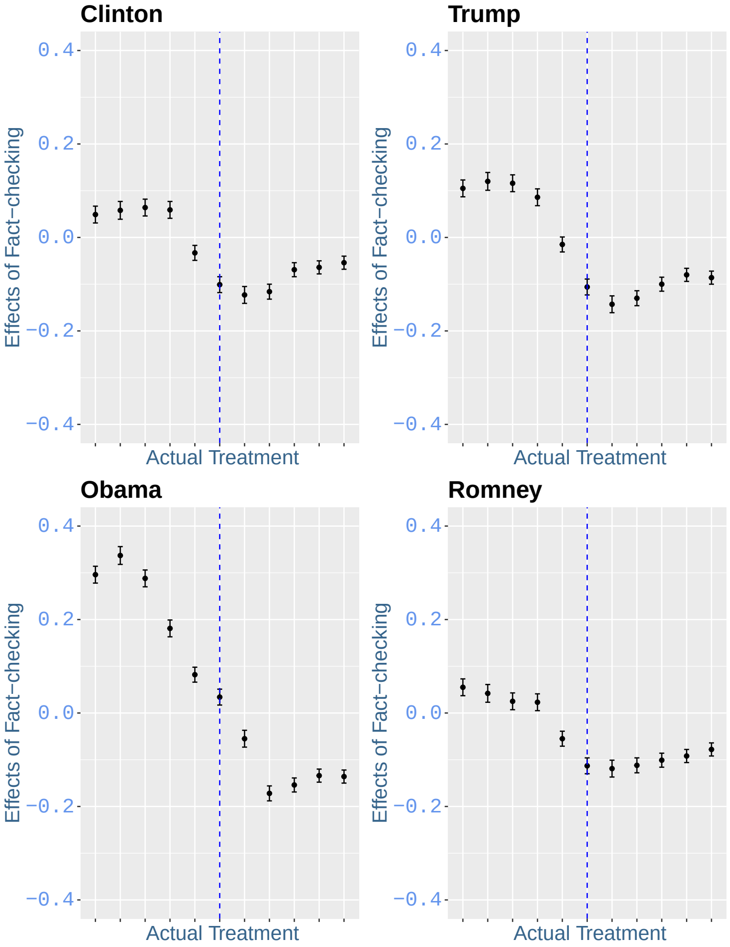
$$Spoken_{it} = \eta_i + \alpha Factcheck_{it} + \beta Length_{it} + \epsilon_{it}$$

Again, I use statement-specific fixed effects and control for speech length. Standard errors are clustered at the statement level.

According to the “Natural decrease over time” hypothesis, candidates gradually stop making certain claims after a certain point, regardless of whether they are fact-checked or not. Thus, I expect to observe the following: 1) 10 coefficient estimates obtained from each round of analyses with the corresponding placebo date of fact-check should all be negative, because there should be a downward trend in the number of statements made after any randomly chosen cutoff date and 2) The magnitude of the negative coefficient on an actual treatment will not necessarily be the largest or among the largest, because fact-check should not affect the probability that a candidate repeats a false-rated claim.

In Figure 5, the 10 points on a plot represent the coefficients obtained from running 10 rounds of fixed effects regression, each with a different placebo date. For example, suppose Statement A was fact-checked on September 19th, 2016. Then, the vertical dashed line in the middle indicates a coefficient from a regression using 9/19/2016 as a treatment date for assigning speeches into either a treatment or control group. Each of the 5 points that lie on the left side of the dashed line represents a coefficient from a regression using 9/14/2016, 9/15/2016, 9/16/2016, 9/17/2016,

Figure 5. Coefficients for Interrupted Time Series Design with Placebo Fact-check Dates Figure 5 displays coefficients obtained from each round of 10 placebo analyses. The vertical dashed line in the middle indicates a coefficient for the actual (non-placebo) date of fact-check. For all candidates, the majority of coefficient estimates that are on the left side of the actual treatment are positive and the magnitude of the negative coefficient on an actual treatment are among the largest.



and 9/18/2016, respectively, as a placebo treatment date. Likewise, each of the 5 points that lie on the right side of the dashed line represents a coefficient from a regression using 9/20/2016, 9/21/2016, 9/22/2016, 9/23/2016, and 9/24/2016, respectively, as a placebo treatment date (See Appendix E for estimates of coefficients and standard errors for each placebo date).

As shown in Figure 5, for all four candidates, the majority of coefficient estimates that are on the left side of the actual treatment are positive, which implies that a downward trend is not present when a cut-off point is randomly chosen on dates other than the actual date of fact-check. The sign of the coefficients turns negative near the actual treatment date. However, the magnitude decreases as the points move farther away from the actual date of fact-check. The magnitude of the negative coefficients on or near the actual treatment date are among the largest. This suggests that the “natural decrease over time” hypothesis alone fails to explain the post-factcheck decrease in the number of false statements.

Testing for Topic Change

Using an unsupervised text classification method, I evaluate the “Topic Switch” hypothesis, according to which a decrease in the percentage of false statements after the date of fact-check may be an artifact of candidates’ decision to move on to a new agenda independent of fact-check. If the “Topic Switch” hypothesis were true, I expect to observe a significant change in speech content after the fact-check.

Of all fact-checks that occurred between August 1st and early November of the election year, about 45 percent of fact-checks are conducted immediately after the three presidential debates. If candidates decide to switch to a new agenda after the debate, they would not repeat many of the claims that were made before the debate. Then, because roughly 45 percent of fact-checks occur immediately after the debate, it would seem as if the decline in the percentage of false-rated claims were caused by fact-checking, when it could have instead been a result of candidates’ decision to move on to a new set of topics after the debate for reasons unrelated to the fact-checking itself.

To find out if there was a significant switch in topics after the debate, I compare the content of speeches before and after each presidential debate. I apply latent Dirichlet allocation (LDA) to classify 30 speeches per candidate. These 30 speeches consist of 15 (3 sets of 5) speeches that were made before each of three presidential debates and 15 (3 sets of 5) speeches that were made after the debates (See Appendix F for the LDA classification results). I assume that the collection of speech transcripts is driven by 5 topics, a number chosen after assessing the substantive fit within and among the clusters. Since different words in a document may be generated from different topics, each document is represented as a mixture of different proportions of various underlying topics. I then assigned the topic with the maximum proportion to each document (Blei et al. 2003). Due to a small sample size (30 speeches for each candidate), the LDA classification may be a bit noisy. Yet, the method still offers a useful comparison of what the topic distribution looks like before and after each presidential debate.

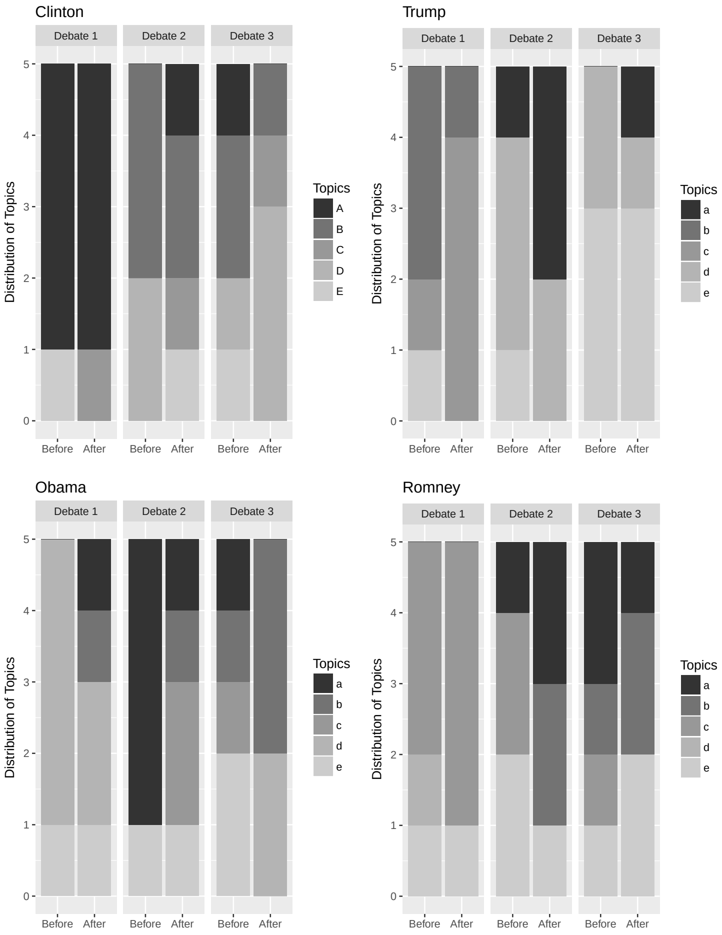
Figure 6 shows the distribution of LDA-classified topics for 5 speeches that were made before and after each round of three Presidential Debates for each candidate. For Clinton, Obama, and Romney, because the second and the third presidential debates are only 6 days apart in 2012 and 10 days apart in 2016, 4 out of 5 “After” speeches for Debate 2 are exactly the same as the first 4 “Before” speeches for Debate 3. Thus, the topic distribution of speeches that were made after Debate 2 looks almost identical to the topic distribution of speeches made before Debate 3. Although the topic distribution among speeches made before the debate is slightly different from the topic distribution among speeches made after the debate, candidates tend not to dramatically shift their agenda so quickly immediately once the debate ends. Instead, they continue to cover a similar set of topics after the debate.

Next, I ran Fisher’s exact test to evaluate if there was a significant change in the assignment of topics after the debate. Fisher’s exact test confirms that for all 4 candidates, the differences in topic distributions after each debate are not statistically significant (See Appendix F for information on p-values from Fisher’s exact test for each candidate). The results show little support for “Topic Switch” hypothesis, implying that topic change is not the main reason why candidates stop repeating many of the debunked claims after the fact-check.

Does the Type of Campaign Event Affect the Effects of Fact-checking?

Given that fact-checking has a pretty significant deterrence effect, does the effect size vary depending on the context of a speech? Presidential debates typically

Figure 6. Distribution of Topics Before vs. After Presidential Debates Figure 6 shows the distribution of topics for 5 speeches that were made before and after each round of three Presidential Debates for each candidate.



get more media attention than any other types of campaign events. In 2016, on C-SPAN, the average number of views for a presidential debate was 48,181, significantly higher than that for smaller campaign events, for which the average number of views was around 3,000. Presidential debates are also the busiest time of year for fact-checkers. 45 percent of fact-checks from August to early November are evaluations of statements that were made during a presidential debate. Fact-checkers are also way more active on social media around the time of presidential debates. For example, Politifact tweets twice as often during and immediately after a presidential debate.

Knowing that their words are likely to receive more media scrutiny around the time of debates, candidates' reaction to negative fact-checks may be different during this time. First, fact-checkers' negative ratings for candidates' statements that

Table 5 Diff-in-Diff Estimates for Debate vs. Non-Debate Claims (Columns 1-3)& Diff-in-Diff Estimates for “Debate Post Factcheck” vs. “No Debate Post Factcheck” Claims (Columns 4-6) Although statistically insignificant, the difference-in-differences estimates in columns 1-3 are negative, which suggests that although the effects may be small, candidates are slightly more likely to stop repeating a debunked claim that is made during a presidential debate and fact-checked immediately after, compared to claims that are made and fact-checked during less salient campaign events. The diff-in-diff results in columns 4-6 imply that there is very little systematic difference in the effects of fact-checking among “debate post fact-check” and “non-debate post fact-check” statements.

	Debate vs. Non-Debate			“Debate Post Factcheck” vs. “No Debate Post Factcheck”		
	All	2016	2012	All	2016	2012
Factcheck	-0.033 (0.026)	−0.009 (0.030)	−0.180 (0.067)	0.001 (0.037)	0.025 (0.042)	−0.076 (0.081)
Statement F.E.	✓	✓	✓	✓	✓	✓
Speech F.E.	✓	✓	✓	✓	✓	✓
Observations	2230	1860	370	2230	1860	370

were made during a presidential debate may be more effective (compared to fact-checks conducted during less salient campaign events) in preventing candidates from repeating these debunked statements in the future. Second, because fact-checkers tend to be more active in scrutinizing candidates’ words around the time of presidential debates, candidates are more likely to be called out if they repeat a claim that has already been debunked by fact-checkers in the past (before the debate). Then, candidates may be less likely to repeat a debunked claim around the time of presidential debates. I evaluate each of these predictions using two different types of difference-in-differences design.

To test if the effectiveness of fact-checks depends on the saliency of campaign events in which a fact-checked statement is made, I run a difference-in-differences test for “debate” statements (statements that are made during presidential debates) and “non-debate” statements (statements that are made in less salient campaign events). Columns 1-3 of Table 5 display difference-in-differences estimates for all candidates, candidates for 2016, and candidates for 2012, respectively. The difference-in-differences estimate is statistically insignificant and smaller in 2016 than in 2012, but the coefficients are negative for both election years. Candidates are slightly more likely to stop repeating a debunked claim that is made during a presidential debate and fact-checked immediately after, compared to claims that are made and fact-checked during less salient campaign events.

Next, I evaluate if candidates are less likely to repeat a debunked claim around the time of presidential debates possibly due to increased fear of being called out by a fact-checker since fact-checkers are particularly active during this time. I identify statements for which post fact-check speeches include a presidential debate. For instance, since the three presidential debates in 2016 are held on September 26, October 9, and October 19, any statement for which the 5 post fact-check speeches include one of these dates are chosen. As an example, if a statement was fact-checked on October 7, the second presidential debate (October 9) is included as one of the post fact-check speeches. Then, I compare the effects of fact-checking on these statements vs. statements for which post fact-check speeches do not include a presidential debate. For convenience, these statements are each labeled "debate post fact-check" and "non-debate post fact-check" statements. Columns 4-6 of Table 5 show difference-in-differences all candidates, candidates for 2016, and candidates for 2012, respectively. The difference-in-differences coefficient is slightly positive in 2016 and is negative for 2012. Possibly due to a small sample size, the estimates are quite small and statistically insignificant for both years. The results imply very little systematic difference in the effects of fact-checking among "debate post fact-check" and "non-debate post fact-check" statements.

It appears that the saliency of campaign events does little to reinforce the effects of fact-checking. Candidates are only slightly more likely to stop repeating a debunked claim that is made during a presidential debate, compared to claims that are made during less salient campaign events. Moreover, despite the higher likelihood of getting called out if they repeat a claim that has already been debunked by fact-checkers in the past, candidates do not seem to take any more caution during the debates.

Discussion & Conclusion

Journalists now regularly trumpet fact-checking as an important facet of watchdog journalism to hold public figures accountable for what they say (Graves 2016). With the rise of online elite fact-checkers and a surge in fact-checking by newspapers and TV stations, fact-checking has become a popular form of media scrutiny especially during presidential election campaigns.

Yet, despite the popularity, the real-world effects of fact-checking during presidential campaigns have received little scholarly attention. Using a rigorous research design, this paper finds that among presidential candidates for the 2012 and 2016 elections, a fact-checking agency deeming a statement false is associated with a 9.3 percentage point decrease in the probability that the statement is repeated in the future.

Then, why do candidates make false claims in the first place if they are going to wind up pulling them back? One simple explanation may be that candidates' lies were genuine mistakes and they correct them once the claims are debunked by fact-checkers. Alternatively, there may be instances where candidates make misleading claims even when they know that such claims may not stand up to the scrutiny of fact-checkers. This may be because candidates believe that a negative fact-check would not be detrimental enough to affect their chances of winning. They may be confident that a solid group of loyal voters will show unwavering support even in the face of fact-checkers' accusations. According to Chuck Todd on NBC's Meet the Press,[13] Trump supporters care very little even if he strays from the facts. Todd also points out that "despite their problems with the truth, Trump and Clinton remain their parties' frontrunners".

Moreover, even if candidates are caught lying, they may feel that they can easily find ways to downplay its seriousness by emphasizing that their opponents engage in a worse form of"truth-bending". Or, they can defend themselves by arguing that fact-checkers are biased and hence not to be trusted. For instance, fact-checkers are

[13] https://www.nbcnews.com/meet-the-press/meet-press-november-29-2015-n470871

often accused by Republicans of letting Clinton “slide” with falsities. During both 2012 and 2016 presidential elections, fact-checkers were mocked for trying to act like “mighty jurists” even when they were proven to be factually incorrect in some cases (See Appendix H for examples of politicians’ accusation of fact-checkers for being biased.).

Also, given that fact-checkers may not always be infallible in the process of selecting and evaluating political claims, we may expect that fact-checkers’ evaluations may not carry as much weight (Lim 2018). Yet, this study finds that contrary to stories presented above, a negative fact-check rating was associated with a decrease in the probability that candidates would repeat false-rated in the future. Here are my speculations as to why fact-checking seems to be associated with a change in candidates’ behavior:

First, fact-checkers track politicians who repeat claims that have already been found to be false. Once candidates are caught repeating a debunked claim, fact-checkers report them in a special column dedicated to repeated false claims, such as “Recidivism Watch” on Fact Checker and “Groundhog Friday” on FactCheck.org. When candidates are repeatedly accused of lying and refusing to correct their claims even after learning that these claims have been debunked by fact-checkers, they may lose support from voters. In addition, candidates might worry that negative ratings from fact-checkers may cause donors or other political elites to withdraw their endorsements.

Second, personality may factor into candidates’ decision to pull back claims that have been debunked by fact-checkers. Candidates may be afraid that being called a liar may hurt their reputation. For example, on CNN’s State of the Union, Rawlings-Blake, secretary of the Democratic National Committee, notes that Clinton, for whom fact-checking seemed to have the biggest effect, “cares about [her] reputation and character” and thus takes it very personally when she is called a liar.[14]

Third, another possible explanation may be that sometimes, candidates might

[14] http://www.cnn.com/TRANSCRIPTS/1608/07/sotu.01.html

have been genuinely unaware that they were lying. In this case, as mentioned above, candidates may willingly correct themselves once they learn that fact-checkers have found their statements to be false.

Across research designs, the findings in this paper suggest that contrary to anecdotal evidence, negative fact-checks is associated with a reduction in the number of false-rated claims in speeches made after the fact-check. The results imply that fact-checking may have an impact beyond merely being used as a rhetorical tool by candidates in their campaigns. According to Bill Adair, the founder of Politifact, these findings are consistent with what he has heard from campaign officials and party leaders, who have said that “they do indeed care about fact-checking” (Bill Adair, personal communication, May 3, 2018). In showing that news organizations may affect candidate behavior, this study contributes to the literature on how news media can hold politicians accountable.

References

Ansolabehere, Stephen, Erik C. Snowberg, and James. M. Snyder. 2006. “Television and the incumbency advantage in US elections.” Legislative Studies Quarterly 31:4: 469–490.

Barrett, Paul. 2001. “Interrater reliability: Definitions, formulae, and worked examples.” Retrieved March 7, 2018 (http://www.pbarrett.net/presentations/rater.pdf)

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent dirichlet allocation.” Journal of machine Learning Research 3 (January): 993–1022.

Cappella, Joseph N. and Kathleen Hall Jamieson. 1994. “Broadcast adwatch effects: A field experiment.” Communication Research 21:3: 342–365.

Carr, David. 2012. “A last fact check: It didn’t work.” Retrieved February 20, 2018 (https://mediadecoder.blogs.nytimes.com/2012/11/06/a-last-fact-check-it-didnt-work/)

Gottfried, Jeffrey A., Bruce W. Hardy, Kenneth M. Winneg, and Kathleen Hall Jamieson. 2013. “Did fact checking matter in the 2012 presidential campaign?” American Behavioral Scientist 57:11: 1558–1567.

Graves, Lucas. 2016. Deciding what’s true: The rise of political fact-checking in American journalism. New York: Columbia University Press.

Guess, Andrew, Brendan Nyhan, and Jason Reifler. 2018. “Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign.” The Knight Foundation.

Hassan, Naeemul, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. “Toward automated factchecking: Detecting check-worthy factual claims by claimbuster.” In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 1803–1812.

Lim, Chloe. 2018. “Checking How Fact-checkers Fact-check.” Research & Politics July-September 2018: 1–7.

Marietta, Morgan, David C. Barker, and Todd Bowser. 2015. “Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities?” The Forum 13: 577–596.

Min, Young. 2002. “Intertwining of campaign news and advertising: The content and electoral effects of newspaper ad watches.” Journalism & Mass Communication Quarterly 79:4: 927–944.

Mummolo, Jonathan. 2018. "Modern police tactics, police-citizen interactions, and the prospects for reform." The Journal of Politics 80, no. 1: 1-15.

Nyhan, Brendan and Jason Reifler. 2010. “When corrections fail: The persistence of political misperceptions.” Political Behavior 32:2: 303-330.

Nyhan, Brendan and Jason Reifler. 2015. “The effect of fact-checking on elites: A field experiment on us state legislators.” American Journal of Political Science 59:3: 628–640.

O’Sullivan, Patrick B. and Seth Geiger. 1995. “Does the watchdog bite? newspaper ad watch articles and political attack ads.” Journalism & Mass Communication Quarterly 72:4: 771–785.

Shao, Chengcheng, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2018. "Anatomy of an online misinformation network." PloS one 13, no. 4: e0196087.

Shin, Jieun, Lian Jian, Kevin Driscoll, and François Bar. 2017. "Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction." New Media & Society 19, no. 8: 1214-1235.

Snyder Jr, James M. and David Strömberg. 2010. “Press Coverage and Political Accountability.” Journal of Political Economy 118:2: 355–408.

Stencel, Mark. 2015. “Politicians modify words, prepare evidence to satisfy factcheckers.” Retrieved February 10, 2018 (https://www.americanpressinstitute.org/publications/reports/survey-research/politicians-modify-words-prepare-evidence-satisfy-fact-checkers/)

Wintersieck, Amanda L. 2017. “Debating the truth: The impact of fact-checking during electoral debates.” American Politics Research 45:2: 304–331.

Wood, Thomas, and Ethan Porter. 2016. "The elusive backfire effect: Mass attitudes’ steadfast factual adherence." Political Behavior 1-29.

Ideology and News Content in Contested U.S. House Primaries

August 28, 2020

Abstract

Pundits and scholars often claim that congressional primary elections favor extremist candidates, but the mechanisms by which primary voters might learn about candidate platforms are not well understood. In this paper, we collect a new dataset of roughly 16,000 local newspaper articles matched to candidates in contested U.S. House primary races from 1998 to 2012. Using supervised machine learning, we classify these articles into political topics. On average, we find little coverage of candidate platforms. However, we also find that the advantage of extremist candidates in House primaries—measured using the campaign contributions they receive—is concentrated in elections with low levels of newspaper coverage. Where newspaper coverage is higher, there is more coverage of candidate platforms, and extremist candidates do worse. The results suggest that the advantage of more extreme candidates in contested House primaries may be the result of information failures and not just the preferences of primary electorates, and that extremist candidates may do increasingly well as local newspaper coverage continues to decline.

1 Introduction

Polarization in U.S. legislatures is at all-time highs (e.g., McCarty, Poole, and Rosenthal 2006), leading scholars and pundits to search for its roots in electoral politics. A popular claim is that congressional primary electorates prefer more extreme candidates, which in turn causes legislative polarization (see, e.g., Owen and Grofman 2006; Pildes 2011, but also see Boatright 2013; Hirano et al. 2010; McGhee et al. 2014). A policy brief from Brookings, for example, writes that "The electorates in [primaries] tend to be small...and often unrepresentative. Hence, candidates are frequently forced to protect their flanks by moving away from the center."[1] Existing research does suggest that congressional primary electorates prefer more extreme candidates, on average (Brady, Han, and Pope 2007; Hall and Snyder 2015*a*; Thomsen 2018), but the magnitude of the advantage is modest. At the same time, it is still an open question whether primary electorates are unrepresentative of voters more generally or not (Hill and Tausanovitch 2017; Sides et al. 2017). Whatever the truth, a missing link in any account of whether or how primary electorates support more extreme candidates is information; if primary voters intentionally support more extreme candidates because of their platforms, then at least some pivotal subset of voters must have information about the platforms of primary candidates. Little research has directly investigated what information is available to primary voters in real elections.

To accomplish this goal, we collect a new dataset from online sources containing the headlines and summaries of approximately 16,000 local newspaper articles about primary candidates in U.S. House races over the time period 1998 to 2012. Based on a careful reading of a random sample of several thousand of these headlines and summaries, we use supervised machine learning to sort the articles into six mutually exclusive categories. After validating the classifications using third-party hand codings and correlations with real-world events, we find that, on average, local news provides voters with little information about their primary candidates' platforms. The average candidate in a contested House primary

[1] https://www.brookings.edu/research/thinking-about-political-polarization/

is mentioned in only 3.3 articles in total, and we estimate that 75% of these articles provide what we call basic campaign coverage—articles about the bare facts of the race, like who is running and who has dropped out, that convey no ideological information at all. More than three quarters of the races in our sample appear to have no news articles covering candidates' stated position at all, while another three quarters have no articles about endorsements, and more than half have no articles about platforms or endorsements, combined.

This does not mean that news coverage is not informative, however. We also show that the share of news coverage in a primary race is a useful predictor of who will win the race. Even if primary news coverage is often short on details, newsworthiness itself is an important leading indicator of electoral outcomes.

Finally, we show that newspaper coverage appears to help more moderate candidates—contrary to the claim that primary voters prefer more extreme candidates. Using estimates of candidate ideology estimated from campaign contributions, we find that primary electorates' preference for more extreme candidates is concentrated in low news-coverage areas. Where newspaper coverage of the primary election is higher, more moderate candidates receive higher average vote shares than where it is lower. Although our measure of newspaper coverage is not randomly assigned, we do what we can to suggest that the relationship is causal by using a potentially exogenous measure of local news coverage based on congruence between a newspaper's circulation and the local congressional district (Snyder and Stromberg 2010).

Adding to the plausibility of these analyses, we find that, where congruence is higher, there are more news articles about candidate platforms. This suggests that primary voters may support more extreme candidates less when they learn more about candidate positions. It is difficult to square these results with the claim that primary voters genuinely prefer more extreme candidates; if that claim were true, more news coverage of candidate positions should help primary voters to pick out more extreme nominees.

In addition to its relevance for the study of primary elections and polarization, our paper also adds to the literature on voters and information in two main ways. First, by examining news coverage directly, we are able to evaluate the actual information environment that primary electorates face. This is in contrast to survey studies that manipulate the information environment, but at the risk of not reflecting the real-world information environment (e.g., Fowler and Margolis 2014; Riggle 1992). The patterns of news coverage that we document may be useful for future survey-based studies that wish to emulate the real news environment. Second, we are able to study the key mechanism by which aggregate patterns of news coverage—investigated in, for example, Peterson (2017) and Snyder and Stromberg (2010)—actually influence public opinion and electoral choices. The fact that higher congruence areas also see more coverage of candidate platforms suggests that media coverage may influence voter behavior directly. In addition, the fact that news worthiness is itself a predictor of candidate performance suggests that newspaper coverage may help primary voters to vote strategically and avoid wasting votes (e.g., Hall and Snyder 2015*b*).

Taken together, our evidence suggests that the continued decline of local news organizations—documented, for example, in Peterson (2018)—will increase the tendency to nominate more extreme candidates, unless alternative sources of information substitute for the campaign coverage that local news has historically provided. This is consistent with evidence concerning local television news (Martin and McCrain 2019).

2 What Can Primary Voters Learn From News Coverage?

In this section, we evaluate the content of local newspaper headlines and article summaries related to contested U.S. House primary elections.

2.1 New Data on Newspaper Articles About House Primaries

We collected information on primary news articles from NewsLibrary.com, an online archive of articles published by over 6,000 newspapers from across the United States. It has a wide variety of local newspapers of varying circulations, ranging from large-scale newspapers like *The Sacramento Bee* and *New York Daily News* to small-town daily newspapers such as the *Eaglewood Sun*. NewsLibrary.com was used in Snyder and Stromberg (2010) to collect local news articles about members of the U.S. House.

To query NewsLibrary, we start from a dataset on U.S. House primary elections, originally collected by Ansolabehere et al. (2010) and extended to subsequent years by the same authors. Along with a full range of electoral variables (like vote share), the dataset contains the full name of each candidate. Using NewsLibrary, we collected the headlines and summaries for local newspaper articles for these candidates in all contested U.S. House primary elections from 1998 to 2012.[2] Each search was confined to newspapers that were published in the state in which the candidate ran between January 1st and the primary election date for each election year and state, using the following search terms: [Candidate's Last Name] in Headlines, ["Primary"] in All Text, and ["Candidate"] in All Text. We deleted articles that were published on or after the primary election date (which varies across states) for each corresponding state, because we are focused primarily on information that is available to voters prior to the election day.[3]

To clean the resulting dataset, we manually scrutinized articles for candidates whose last names are among the 100 most common in the 2010 census, to determine whether each article was in fact about a House primary candidate. We likewise manually checked articles where the lead paragraph did not include the first name of the candidate. By reading all of these articles, we excluded those written about a different person with the same last name running for different office (e.g., County Commissioner, District Attorney, etc.) or about the

[2] The archive does not grant open access to full article content, but it does allow us to view headlines and summaries that provide relatively detailed information about the article.

[3] We obtained data on primary election dates from the FEC website.

same person running for another office in the same election year, after having resigned from the House race.

We also use the FEC IDs from the election dataset to merge candidates with their ideological scalings originally developed in Hall and Snyder (2015*a*), and extended to later years in Hall and Thompson (2018). These scalings impute ideological positions for candidates who've never held office in a two-step process. First, donors are scaled based on the DW-NOMINATE scores of incumbents that they donate to—so, for example, a donor who donates to candidates who have far-right Nominate scores is imputed to be a far-right donor. Second, candidates are scaled based on the donors from whom they received contributions—so, for example, a candidate who receives donations from donors that support far-right incumbents is estimated to be a far-right candidate. The resulting scalings correlate well with DW-NOMINATE, even within-party. For further discussion of the validity of these scalings, see Hall and Thompson (2018).

The final dataset includes 2,448 candidates and 15,801 articles published in 2,039 local newspapers.

2.2 Classifying Primary News Coverage

Our goal is to understand the content of newspaper coverage about U.S. House primaries, and to evaluate whether it offers significant information about candidate platforms. Before applying any methods, we read 5,000 article headlines and summaries ourselves. Based on our reading, we defined six categories of news coverage: Campaign Coverage, Endorsements, Candidate Biographies, Money, Platform, and Scandal. The categories are largely self explanatory. Campaign Coverage articles discuss the bare facts of the race, such as who is running and who has dropped out. Articles in the Endorsements category report endorsements that candidates have received. Articles in the Candidate Biographies category provide specific information on candidates' backgrounds, like their professions, any previous political offices they have held, their age, and so forth. Articles in the Money category are focused on information about candidate fundraising. Articles in the Platform category, which we are particularly interested in, report specific policy positions or ideological views that candidates have offered. Finally, articles in the Scandal category focus on potential scandals related to a candidate in the race. Because we do not have access to full article content, it is possible that an article classified in one category based on its headline and summary could contain paragraphs that would fit into other categories; however, when we compared full content for articles we were able to find on other websites online, we rarely found this to be the case (most articles are quite short, and the summaries generally indicate the full scope of their content).

Appendix A offers more details on the categorization scheme used to classify the news articles, and offers specific examples of articles coded into each category. We automatically coded any article including the word stem "endors" as Endorsement and assigned 4,828 articles to one of the remaining five categories manually.

We then used these 4,828 researcher-coded entries as a training set for a variety of supervised machine learning procedures (see for example Grimmer and Stewart 2013; Lucas et al. 2015). Specifically, we compared the performance of eight different classifiers: Support Vector Machine, Maximum Entropy, Supervised Latent Dirichlet Allocation, Boosting, Bagging, Random Forest, Neural Network, and Tree (we did not include the Endorsement category articles since their coding is deterministic.) Of the 4,828 entries, 2,414 were randomly chosen to train each model. We measured how well each model performed on the entries that were not used to train the model by comparing the classification results against the researcher codings.

Table A.1 shows precision, recall, and f-scores for each algorithm. In the context of our research, precision is the proportion of news articles correctly classified as Category A out of the total number of articles classified as Category A by the algorithm. Recall is the proportion of news articles correctly classified as Category A out of all true Category A articles (i.e., all articles assigned to Category A by the researcher). Recall is a function of both true positives

(Category A articles correctly assigned to Category A) and false negatives (Category A articles incorrectly assigned to other categories). F-scores are a weighted average of both precision and recall.

We choose to focus on SVM, since it has the highest F-score, at 0.734. Having chosen SVM, we trained it on the entire training set of researcher-coded entries. We then applied it to the rest of the newspaper articles which were not classified by the researcher, providing us with topic classifications for each article in our sample.

2.3 Validating Our Primary Election Newspaper Coverage Classifier

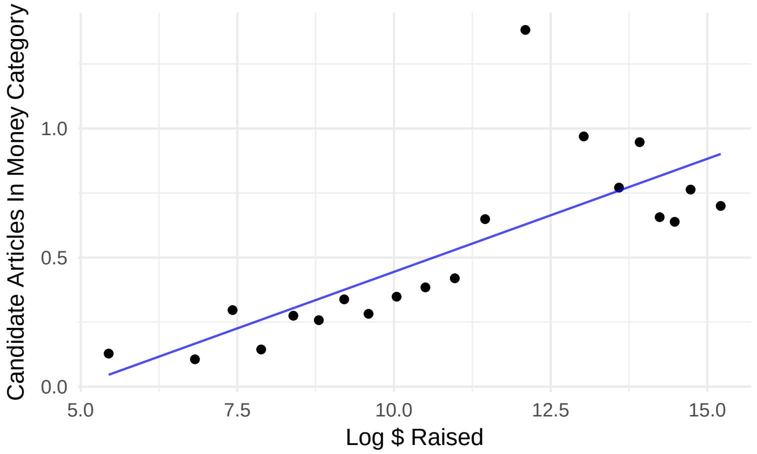
Before using the SVM classifications to analyze news content, we validate our approach in several ways. First, and most importantly, we asked an independent coder to categorize a random sample of 1,000 articles from the test set. The intercoder reliability rate between our codings and the codings of the independent coder was 0.77, meeting the standard of intercoder reliability that academics commonly apply when evaluating hand-coded data (Barrett 2001).

Second, we correlate the amount of money each candidate in our dataset raises with the number of articles about that candidate that are classified in the 'Money' category. Figure 1 presents a binscatter showing the strong correlation between the two variables. Candidates who raise more money also have more articles classified into the Money category, suggesting that, at the very least, the Money topic categorization is meaningful.

Third, we investigate which specific words are most predictive of each categorization. To do so, we create a document-term matrix (DTM) from the article summaries, and we combine this matrix with the SVM categorizations for each article.[4] We construct a dummy variable for whether each article is a member of each of the six categories, and we run ridge regressions

[4] We construct the DTM using the create_matrix() function in the RTextTools package, using options to remove numbers, stem words, and remove stopwords. We set the removeSparseTerms option to 0.99.

Figure 1 – Validating Article Classification Using Fundraising Data. Candidates who raise more money have more articles classified in the 'Money' Category.



Note: Points are averages in equal-sample-sized bins of Log $ Raised. Regression line fit to underlying data. Generated using `binscatter` in Stata.

Table 1 – Informative Words by Category. Presents the five words most predictive of an article being classified as each of the six mutually exclusive topics.

Campaign Info	debate,district,candidate,percent,rep
Bio	series,education,university,candidate,college
Endorsement	endorsed,endorsement,endorsements,endorse,endorses
Money	money,raised,fundraising,fund,million
Platform	tax,jobs,health,federal,abortion
Scandal	accused,court,campaign,federal,commission

predicting each membership dummy using the features from the DTM. Table 1 presents the five words most predictive of each category—i.e., the five words with the largest coefficients in the ridge regression. As the table shows, the predictive words are highly sensible for all six categories. The campaign info category features generic words about candidates; the bio category features words about education, because candidate biographies almost always discuss candidates' educational backgrounds; the endorsement category by construction is based off of the word stem "endors"; the money category features words about money; the platform category features key policy words, like tax, health, and abortion; and the scandal category features words like accused and court, as would be expected.

Table 2 – Summary Statistics. Number of articles mentioning a candidate, by topic. Unit of observation is a candidate-year.

	Mean	SD	Min	Max	N
Total Articles Mentioning Candidate	3.33	7.26	0.00	156.00	4,747
Campaign Articles	2.08	4.87	0.00	118.00	4,747
Money Articles	0.31	1.29	0.00	36.00	4,747
Platform Articles	0.31	1.16	0.00	40.00	4,747
Scandal Articles	0.10	0.82	0.00	39.00	4,747
Biographical Articles	0.07	0.50	0.00	12.00	4,747
Endorsement Articles	0.45	1.82	0.00	60.00	4,747

3 Evaluating Primary Election Newspaper Coverage

We now turn to studying what features of primary elections local newspapers cover, using the dataset of newspaper coverage generated using the machine-learning procedure we just described.

3.1 How Much Newspaper Coverage of Primaries Is There?

We begin by offering some simple facts about the quantity of newspaper coverage in U.S. House primaries, focusing on the set of races for which we have access to newspaper data. Table 2 offers a summary of the data, where the unit of observation is a candidate in a given year's primary election in a given congressional district, for the set of contested U.S. House primaries. As the first row shows, on average, a candidate is mentioned in 3.33 newspaper articles over the course of the primary election—a modest level of overall coverage. This is consistent with what we know about House primaries from existing accounts; they are low information affairs with little coverage or polling.

The next rows break the coverage down by topic. As the second row shows, the majority of articles are classified as campaign coverage. As a reminder, this category includes articles that cover the basic nuts and bolts of campaigns—who is running, who has dropped out, and so forth. The remaining categories are all much rarer.

Figure 2 – Types of Primary Election News Coverage Over Time, Contested U.S. House Primaries, 1998–2012. Campaign coverage dominates newspaper coverage of primary elections.

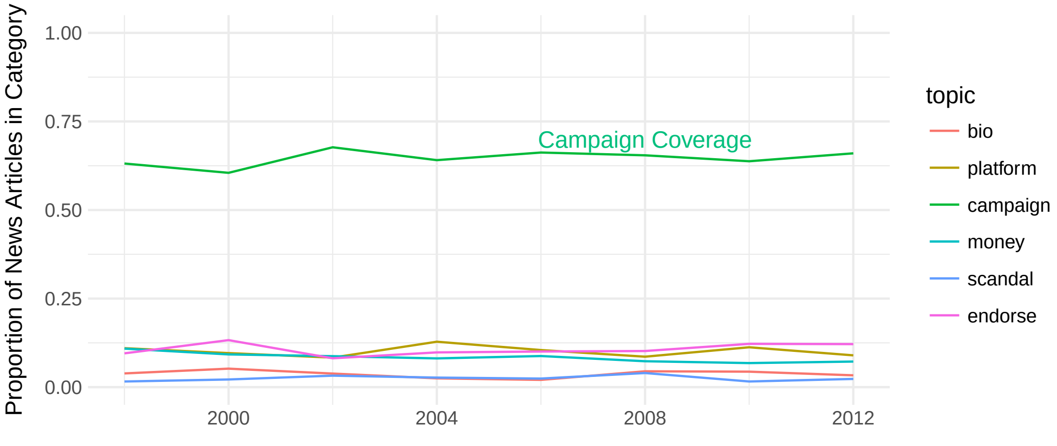


Figure 2 shows the proportion of each category in primary election news coverage over time. Campaign coverage consistently dominates newspaper coverage of House primaries. The other five categories are all much less common. Articles about candidate platforms are rare, never accounting for more than 8% of all articles in any given year. Articles containing news about endorsements are also relatively rare, accounting for roughly 1 out of every 10 articles in the sample.

3.2 Platform and Endorsement Coverage is Relatively Rare

Although aggregate rates of platform and endorsement articles are low, it is possible that only a minimal number of such articles are necessary in order to inform attentive readers. Accordingly, we also count the number of races in which we find zero articles about platforms and endorsements, respectively. Table 3 shows the number of candidate years that receive some or no platform coverage and some or no endorsement coverage. As the upper left cell shows, in the majority of cases, 3,446 in all, candidates receive no newspaper coverage in the platform or endorsement categories. 324 candidates have at least one article mentioning an endorsement but have no articles about their platforms, while 312 have at least one article discussing their platforms but no articles about their endorsements. In only 77 cases do we

Table 3 – Rates of Platform and Endorsement Coverage. Presents the number of candidates who have, or do not have, any coverage of platforms and/or endorsements.

	No Platform Articles	One or More Platform Articles
No Endorsement Articles	3446	312
One or More Endorsement Articles	324	77

Unit of observation is a candidate-election.

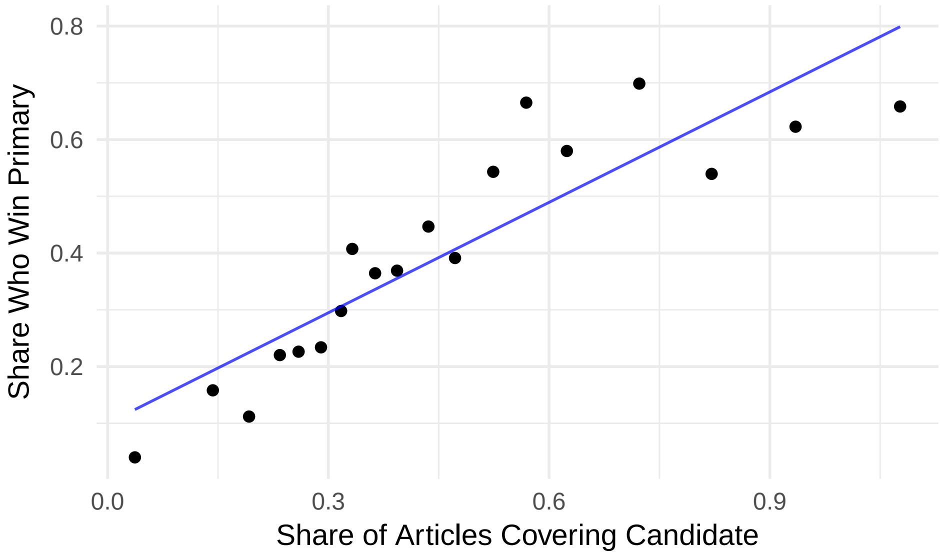
find both types of articles. Overall, only 17% of all candidate-years have any newspaper articles categorized as either platform or endorsement articles.

3.3 Newspaper Coverage Predicts Candidate Success

Although on average newspaper coverage appears to offer voters little direct information about candidate positions, it would be a mistake to conclude that it is altogether uninformative. In fact, receiving newspaper coverage is a strong predictor of candidate success in primary elections, which means that newsworthiness can be an informative signal for voters (note that there does not need to be a causal link between newspaper coverage and electoral performance for newspaper coverage to be an informative signal). Figure 3 plots average nomination rates against the average share of all articles in a primary that a particular candidate receives, in equal-sample-sized bins. As the plot shows, candidates who receive a larger share of newspaper coverage are substantially more likely to win nomination (R^2=0.29).

This relationship may be important because it provides information to voters in the absence of polls. In primaries with three or more candidates, there is a substantial risk of vote wasting (Duverger 1954). Strategically speaking, a voter should only vote for one of the top two candidates in the race, since a vote for any lesser candidate will have no impact on the winner. Hall and Snyder (2015*b*) shows that there is a meaningful amount of vote wasting in House primaries, and that this vote wasting goes down where media coverage is higher. The paper speculates that media information helps voters avoid vote wasting by conveying

Figure 3 – Primary Candidate News Coverage and Electoral Performance; Competitive Primaries in Open-Seat U.S. House Races. The share of news coverage devoted to a candidate predicts the candidate's success in the primary.



Note: Points are averages in equal-sample-sized bins of Log $ Raised. Regression line fit to underlying data. Generated using `binscatter` in Stata.

information about who the favored candidates are. Figure 3 supports this hypothesis; even in the absence of much polling of House elections, newsworthiness itself may help voters to identify who the favored candidates are. If a voter observes a candidate garnering a higher degree of newspaper coverage, she can infer that the candidate is expected to do well in the primary election.

It does not appear that newspaper coverage directly talks about electability; rather, simply the volume of news coverage is informative. To investigate this, we used a set of following words to look for articles that were informative of candidates' electability: "poll", "survey", "result", "percent", "win", "lead"/"led", "financial lead", "raise", "frontrunner", "underdog", "first place", "second place", and "runoff". We then closely read every article that contained any of these words to determine whether a given article actually contained information on the candidate's electability. Most electability articles were about polling results that showed which of the candidates were leading. For those who end up in a run-off, articles on how they performed in primaries were also categorized as electability. We also selected articles containing nouns that were informative of the candidate's ranking, such as "frontrunner" or "first place." Of 3,282 Campaign Coverage articles, only 283 of them (8.6 percent) were informative about candidates' electability.

4 The Preference for Extremists and Newspaper Coverage

Thus far, we have shown that local newspapers offer relatively little coverage of House primaries, that little of this coverage contains information about candidate platforms, but that the level of coverage still provides useful information about candidates' electoral prospects. In this section, we turn to considering the possible effects this coverage has on the decisions that primary electorates make. The evidence we show is consistent with the idea that newspaper coverage informs primary voters about candidates and helps to undo the advantage of extremist candidates.

4.1 Scaling Candidates Using Campaign Contributions

Following Hall and Snyder (2015*a*), for each candidate i in primary election k we calculate her "relative extremism" compared to the other candidates in her primary as

$$\textit{Relative Extremism}_{ik} = |\textit{Cand Ideology}_{ik} - \textit{Most Moderate Ideology}_{k}|, \tag{1}$$

where $\textit{Most Moderate Ideology}_{k}$ is the estimated ideology of the most moderate candidate in primary k— the rightmost candidate in a Democratic primary or the leftmost candidate in a Republican primary. By using the absolute distance from the most moderate candidate, we can pool over Democratic and Republican primaries. Candidates with higher values of Relative Extremism are farther into the wings of their respective parties than are their primary opponents. As this definition hopefully makes clear, the idea of “extremism” here is entirely relative, and does not refer to any specific issue position or require any normative judgment on our part.

4.2 Extremists Outperform Moderates in U.S. House Primaries

We begin by comparing the performance of candidates with varying ideology in U.S. House races. To do so, we replicate and update the regressions estimated in Hall and Snyder (2015*a*). Specifically, we estimate equations of the form

$$\begin{aligned} Y_{ijkt} &= \beta_0 + \beta_1\, \textit{Relative Extremism}_{ijkt} + \sum_i \beta_{2i}\, I(\#\textit{Cands}_{jkt} = i) \\ &\quad + \sum_{i=1}^{3} \beta_{3i}\, (\textit{Share of Donations}_{ijkt})^i + \sum_{i=1}^{3} \beta_{4i}\, (\textit{Share of Donors}_{ijkt})^i + \epsilon_{ijkt}, \qquad (2) \end{aligned}$$

where Y_{ijkt} is the vote share or an indicator for electoral victory for candidate i in party j’s primary in district k at time t.

Following Hall and Snyder (2015*a*), we estimate this equation only using data for primary elections without an incumbent candidate, since incumbent candidates may be systematically more moderate and also more likely to win election. We also only estimate the equation for cases where at least two primary candidates receive scalings, since this is necessary to compute the measure of relative extremism. Finally, we re-scale the relative extremism measure so that it has mean 0 and standard deviation 1.

We also include controls for the share of donations and donors that candidate i receives in race party j’s primary in district k at time t. Following Hall and Snyder (2015*a*), we include each of these controls as a flexible polynomial. The idea is to make comparisons among candidates with different estimated ideological positions but who raise similar amounts of money, to address the concern that candidates might look more extreme because they raise less money.

Table 4 – Extremists Outperform Moderates in U.S. House Primaries. These results are a replication, with more data, of those found in Hall and Snyder (2015*a*).

	Vote Pct	Vote Pct	Win	Win
Rel Extremism	0.17	0.35	0.03	0.03
	(0.13)	(0.13)	(0.00)	(0.00)
Controls	Yes	Yes	Yes	Yes
Controls Polynomial	1	3	1	3
# Cand Fixed Effects	Yes	Yes	Yes	Yes
# Observations	5,836	5,836	5,836	5,836
# Elections	2,974	2,974	2,974	2,974

Vote Pct runs 0-100; Win is an indicator variable for winning the primary. Robust standard errors clustered by election in parentheses. Rel Extremism, defined in text, is standardized to have mean 0 and sd 1. Control variables are candidate's share of total donors in primary and candidate's share of total donations.

Table 4 presents the results, for both vote percentage and win probability, using two different specifications: one where the polynomials for the share of donations and of donors in equation 2 are first order, and one where they are third order (as in the equation above). Across all specifications, we see a positive association between relative extremism and electoral performance. A one standard deviation increase in relative extremism predicts a 0.21 or 0.38 percentage point increase in vote share, and a 3 percentage point increase in win probability. These associations are relatively modest, in size, but they are mostly precisely estimated.

The takeaway from this analysis is that extremist candidates tend to win U.S. House primary elections at somewhat higher rates than more moderate candidates. This relationship is purely descriptive and does not reflect the causal effect of a candidate *choosing* a more moderate or more extreme position. Indeed, there are many reasons extremists might do worse or better than moderates—we are only measuring the overall filtering of the primary candidate pool. Whatever the mechanism, this descriptive relationship tells us what types of ideology are represented in our legislatures. We now investigate whether this filtering looks different in primary elections with more or less media coverage, to see whether more moderate or more extreme candidates do better when newspaper coverage is higher.

4.3 Extremist Advantage Concentrated in Low Newspaper-Coverage Elections

To see whether newspaper coverage changes the relationship between candidate ideology and success in primary elections, we use the Snyder and Stromberg (2010) measure of newspaper congruence. Districts with higher congruence are those where more of the newspaper's circulation is within the district, which leads the newspaper to devote more coverage to the district's member of Congress and its congressional elections. Snyder and Stromberg (2010) shows comprehensive evidence that higher congruence districts receive systematically more coverage of their members of Congress, and that voters in these districts are more informed about their members of Congress.[5] We scale this congruence measure to run from 0, in the least congruent district, to 1, in the most congruent district, and we re-estimate equations like those in Table 4 with an interaction between candidate ideology and congruence included.

The original congruence measure from Snyder and Stromberg (2010) only ran through 2004.[6] To extend it, we computed a new congruence measure using the formula outlined in Snyder and Stromberg (2010) with updated circulation-by-county figures from Alliance for Audited Media (AAM). We obtained county-circulation data for 2010 and 2011 for 286 (out of 2,039) local newspapers across 422 (out of 436) congressional districts in our dataset and back-filled the updated congruence measure for years 2003-2012.

The hope of using congruence is that it is an exogenous measure of newspaper coverage. Because congruence depends on the historical dispersion of newspapers and of readers, it may have little or nothing to do with the electoral features of present-day districts. However, we

[5] Many papers have used the congruence measure subsequently. As of this writing, Snyder and Stromberg (2010) has been cited nearly 500 times, according to Google scholar.

[6] Snyder and Stromberg (2010) analyze data from 1982 to 2004. They interpolated congruence data for the years 1983-1990, for which they did not have county-circulation data.

Table 5 – Extremist Primary Advantage Concentrated in Low News Coverage Elections.

	Vote Pct	Vote Pct	Vote Pct	Vote Pct
Rel Extremism	0.66	0.85	0.62	0.81
	(0.27)	(0.27)	(0.27)	(0.27)
Rel Extremism × Congruence	-1.16	-1.21	-1.02	-1.07
	(0.51)	(0.51)	(0.52)	(0.52)
Congruence	1.02	1.02	1.30	1.23
	(0.62)	(0.63)	(0.81)	(0.81)
Candidate Controls	Yes	Yes	Yes	Yes
Cand Controls Polynomial	1	3	1	3
District Controls	No	No	Yes	Yes
# Cand Fixed Effects	Yes	Yes	Yes	Yes
# Observations	4,585	4,585	4,585	4,585
# Elections	2,333	2,333	2,333	2,333

Vote Pct runs 0-100. Robust standard errors clustered by election in parentheses. Rel Extremism, defined in the text, is standardized to have mean 0 and sd 1; Congruence, also defined in the text, runs from 0 to 1, min to max. Candidate control variables are candidate's share of total donors in primary and candidate's share of total donations. District control variable are listed in text.

know that more congruent districts tend to be more rural, because urban areas have many districts served by a small number of large newspapers, and we might suspect there are other differences correlated with urban and rural areas. To account for this possibility, we follow Snyder and Stromberg (2010) and also estimate these regressions including a full set of variables about the districts as controls. Specifically, the controls are: percent urban in district; indicators for percent urban quintile; population density; indicators for density quintile; the number of congressional districts per city; log median income; percent senior citizens; percent military; percent farmer; percent foreign; and percent blue collar.

Table 5 presents the results. The pattern seems clear; the advantage of extremist candidates appears to be higher in low congruence areas, where newspaper coverage is more scant, and lower in more congruent places.[7] It is also somewhat encouraging that the coefficient

[7] We have also estimated these results using dummies for quartiles of the congruence variable, to ensure that our results are not driven by the strong assumption of linearity of the interaction of the two continuous variables (e.g., Hainmueller, Mummolo, and Xu 2018). Results are similar in this alternative setup.

estimates on this interaction variable do not change very much based on which controls we include. Moreover, the difference is large enough in magnitude that, in high congruence areas, the relationship inverts and we observe an advantage for more moderate candidates. In a hypothetical race with the highest level of congruence, we estimate that a one standard-deviation increase in relative extremism is associated with, in the smallest estimate (column 3), a 0.4 percentage-point decrease in vote share (0.62-1.02 = -0.4). Although this relationship is not large in magnitude, it is in the opposite direction as conventional wisdom; more informed congressional primary electorates do not appear to favor more extreme primary candidates.

Perhaps because the advantage of extremist primary candidates in general is not large, we do not find a negative interaction of extremism and congruence when we use a binary indicator for victory as the outcome variable. However, the standard errors are very large, so that the confidence interval contains large negative or positive effects. The coarsening of the outcome variable evidently loses too much information for us to offer meaningful estimates on win probability given our sample size and statistical power.

4.4 More Congruent Areas Receive More Platform Coverage

Extremist candidates appear to perform worse in contested House primaries that occur in areas with more local newspaper coverage. Why might this be the case? In this subsection, we explore how the nature of coverage differs in places with more congruent news coverage, and we find that, in high congruence areas, local newspaper articles offer more information about candidate platforms.

Specifically, we estimate equations of the form

$$\#\, \textit{Platform Articles}_{it} = \beta_1 \textit{Congruence}_{it} + \sum_i \beta_{2i} I(\#\textit{Cands}_{it} = i) + X_{it} + \epsilon_{it}, \qquad (3)$$

Table 6 – More Platform Information Where Newspaper Congruence is Higher.

	# of Articles About Platforms	
Congruence	0.32 (0.09)	0.33 (0.10)
Controls	No	Yes
# Cand Fixed Effects	Yes	Yes
# Observations	2,410	2,410
# Elections	718	718

Congruence runs from 0 to 1, min to max. Robust standard errors clustered by election in parentheses.

where X_{it} is an optional vector of control variables. Like before, we use this vector to attempt to control for potential differences between high congruence and low congruence districts.

Table 6 presents the results. As the table shows, higher congruence areas appear to receive more newspaper articles about candidate platforms. The second column presents the estimates with the inclusion of the full suite of district control variables, finding very similar results. The results suggest that extremist candidates may do worse in more congruent areas in part because voters in these areas have more information about candidate platforms, though there are many steps along that causal chain that we cannot observe in our data.

4.5 Newsworthiness More Informative in More Congruent Areas

Another possibility, not mutually exclusive with the increase in platform coverage, is that newsworthiness is also more informative in high-information areas, because there are more articles in general. If more moderate candidates receive a higher proportion of news coverage, then the increased informedness of news coverage in high congruence areas could help to explain the reduced advantage of more extreme candidates in higher congruence areas. To test this, we run regressions predicting vote share and relative extremism, respectively, as

Table 7 – News Coverage is More Informative in Higher Congruence Areas.

			Rel	Rel
	Vote Pct	Vote Pct	Extremism	Extremism
Article Share	22.63	22.43	0.04	0.09
	(2.79)	(2.80)	(0.22)	(0.22)
Article Share ×	14.66	15.07	-0.57	-0.59
Congruence	(5.59)	(5.70)	(0.43)	(0.43)
Congruence	-6.96	-6.07	0.26	0.28
	(2.08)	(2.22)	(0.29)	(0.30)
District Controls	No	Yes	No	Yes
# Cand Fixed Effects	Yes	Yes	Yes	Yes
# Observations	1,512	1,512	1,076	1,076
# Elections	718	718	606	606

Vote Pct runs 0-100, as does Article Share. Robust standard errors clustered by election in parentheses. Rel Extremism and Congruence, both defined in text, are standardized to have mean 0 and sd 1. District control variable are listed in text.

a function of a candidate's article share, interacting article share with congruence as well. Table 7 presents the results.

As the first two columns show, a candidate's share of articles in a primary—that is, her relative newsworthiness—is strongly associated with primary vote share, even in low congruence places (first row). As the second row shows, this relationship is considerably larger in high congruence areas. This is true with or without district controls. This suggests that newsworthiness is an especially good leading indicator of primary success in high-congruence areas.

This relationship may help to explain the diminished advantage of more extreme candidates in high congruence areas if newsworthiness helps primary voters to pick out more moderate candidates. The second two columns suggest this may be the case, but the results are too imprecise to draw any strong conclusions. In these columns, we see that, in low congruence areas, there is no apparent relationship between a candidate's article share and her relative ideological extremism. However, in higher congruence areas, there is a negative though statistically imprecise relationship—that is, candidates in high congruence areas who receive a higher article share appear to be more moderate, on average, consistent with the possibility that newsworthiness is a signal both of electability and lower extremism.

5 Conclusion

The role of primary electorates in the polarization of American politics is much disputed, with scholars debating whether primary electorates are more extreme than other voters, and debating if primary elections encourage polarization or not. In this paper, we have taken a different approach to studying this question. Rather than attempting to measure the issue preferences of primary voters, we have examined how their behavior in real elections varies along with the information environment. Using a new dataset of local newspaper articles in U.S. House primary elections, we have shown that there is relatively little news coverage of primary elections—and of what coverage there is, very little of it concerns candidate platforms.

However, there is important variation in how much information primary voters receive. Where newspaper coverage is higher, more extreme candidates do worse in contested House primaries. Moreover, in these areas, newspaper coverage offers more information about candidate platforms. Together, the evidence suggests that information about candidate platforms may influence the choices that primary voters make, in the direction opposite to what conventional wisdom would predict. Instead of helping primary voters to pick out extreme candidates, newspaper information about candidate platforms may encourage them to pick more moderate nominees.

The decline of local news coverage is an important story in American politics. Although our evidence does not directly estimate the causal effect of the decline in the capacity of local news coverage, it certainly suggests that further declines may lead to an increasing advantage for more extreme primary candidates. At the very least, our results suggest that this is a phenomenon that warrants further study in the future.

References

Ansolabehere, Stephen, John Mark Hansen, Shigeo Hirano, and James M. Snyder Jr. 2010. "More Democracy: The Direct Primary and Competition in US Elections." *Studies in American Political Development* 24(2): 190–205.

Boatright, Robert G. 2013. *Getting Primaried: The Changing Politics of Congressional Primary Challenges.* University of Michigan Press.

Brady, David W., Hahrie Han, and Jeremy C. Pope. 2007. "Primary Elections and Candidate Ideology: Out of Step with the Primary Electorate?" *Legislative Studies Quarterly* 32(1): 79–105.

Duverger, Maurice. 1954. *Political Parties: Their Organization and Activity in the Modern State.* New York: Wiley.

Fowler, Anthony, and Michele Margolis. 2014. "The Political Consequences of Uninformed Voters." *Electoral Studies* 34: 100–110.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267?297.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing. Xu. 2018. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* .

Hall, Andrew B., and Daniel M. Thompson. 2018. "Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in U .S. Elections." *American Political Science Review* .

Hall, Andrew B., and James M. Snyder, Jr. 2015*a*. "Candidate Ideology and Electoral Success." Working Paper.

Hall, Andrew B., and James M. Snyder, Jr. 2015*b*. "Information and Wasted Votes: A Study of U.S. Primary Elections." *Quarterly Journal of Political Science* 10(4): 433–459.

Hill, Seth J., and Chris Tausanovitch. 2017. "Southern Realignment, Party Sorting, and the Polarization of American Primary Electorates, 1958-2012." *Public Choice* .

Hirano, Shigeo, James M. Snyder, Jr., Stephen Ansolabehere, and John Mark Hansen. 2010. "Primary Elections and Partisan Polarization in the US Congress." *Quarterly Journal of Political Science* 5(2): 169–191.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2): 254–277.

Martin, Gregory J, and Joshua McCrain. 2019. "Local News and National Politics." *American Political Science Review* 113(2): 372–384.

McCarty, Nolan M., Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press Cambridge, MA.

McGhee, Eric, Seth Masket, Boris Shor, Steven Rogers, and Nolan McCarty. 2014. "A Primary Cause of Partisanship? Nomination Systems and Legislator Ideology." *American Journal of Political Science* 58(2): 337–351.

Owen, Guillermo, and Bernard Grofman. 2006. "Two-stage Electoral Competition in Two-Party Contests: Persistent Divergence of Party Positions." *Social Choice and Welfare* 26(3): 547–569.

Peterson, Erik. 2017. "The Role of the Information Environment in Partisan Voting." *The Journal of Politics* 79(4): 1191–1204.

Peterson, Erik. 2018. "Paper Cuts: How Reporting Resources Affect Political News Coverage." Working Paper.

Pildes, Richard H. 2011. "Why the Center Does Not Hold: the Causes of Hyperpolarized Democracy in America." *California Law Review* pp. 273–333.

Riggle, Ellen D. 1992. "Cognitive Strategies and Candidate Evaluations." *American Politics Quarterly* 20(2): 227–246.

Sides, John, Chris Tausanovitch, Lynn Vavreck, and Christopher Warshaw. 2017. "On the Representativeness of Primary Electorates." *British Journal of Political Science* .

Snyder, Jr., James M., and David Stromberg. 2010. "Press Coverage and Political Accountability." *Journal of Political Economy* 118(2): 355–408.

Thomsen, Danielle M. 2018. "When Might Moderates Win the Primary?" In *Routledge Handbook of Primary Elections*, ed. Robert G. Boatright. Routledge pp. 226–235.

Online Appendix

Intended for online publication only.

Appendix A. Categorization Scheme

Election news coverage during contested House primaries are categorized into one of six categories. Here, we describe in detail how the categories are defined, along with examples.

1. Campaign Coverage: information on who has entered or dropped out of the race; overview of who is running against whom; a report on election or polling results; prediction on who's going to win based on polling results; information on campaign activities (e.g., ads, commercials, meeting with voters); backing from a former opponent who is no longer running. However, if a candidate receives endorsement or is backed by an interest group with a clear policy or ideological stance, the article is categorized as Platform. Likewise, if an article describes in detail the issues that were discussed in advertisements or during campaign events, it goes under Platform.

- State Senate majority leader Garagiola 6th District candidate: CUMBERLAND Rob Garagiola, Democratic candidate for Marylands 6th Congressional District, knocked on some doors in the Mapleside area Saturday in his quest to win the upcoming primary election and then unseat U.S. Rep. Roscoe Bartlett. I'm getting a lot of positive feedback. Some people are not aware of the primary election which is a month away today. We're taking one step at a time, said Garagiola, the majority leader in Maryland's state Senate and one of four Democratic candidates...
- Reed drops out of 6th District race: Dr. Maureen Reed has dropped out of the 6th District U.S. House contest, saying she's stepping aside in order to focus the race on beating incumbent Rep. Michele Bachmann. "During the past few days, I have come to the conclusion that a prolonged primary fight only assists Michele Bachmann," Reed wrote on her campaign blog. "I feel that it is time for the DFL to unify behind one candidate in this race." Reed, 57 a resident of the Grant, had been the Independence Party candidate for...
- Poll good news for Brown: Campaign Bits By Tom Waring Polls released on Friday by the Center for Opinion Research show Melissa Brown leading a three-way Republican battle in the 13th Congressional District, while the two Democrats are in a close race. Among Republicans likely to vote in the April 27 primary, Brown has 36 percent of the vote. State Rep. Ellen Bard follows with 20 percent. Al Taubenberger trails with 11 percent. "This poll shows we are in excellent position for a victory in April...

2. Candidate Biography: Biographical information (name, age, past experience, etc.) about a candidate; interviews; section in a newspaper that is specifically dedicated for candidate running in a congressional district in which the newspaper is published.

- Christopher Brent Reilly: Age: 50. He lives: York Township, York County. Education: B.A. in government and politics, University of Maryland. Family: Wife, Lisa, and three children, Patrick, William and Claire. Occupation: York County Commissioner. Hobbies: Fishing; reading; and cooking ethnic food. First job: Shoeshine boy at a barbershop. Attribute/ability he will take to Washington: Experience. I have a proven record of fiscal conservatism. I'm a Conservative and I'm...

• Candidate Q&A - Jonathan Paton, candidate for Congressional District 1: Name: Jonathan Paton Age: 41 Occupation: Self, Public Relations Where do you live? Pima County How long have you lived in the area? My entire life. Short description about yourself (200 words) After serving for two years as a Representative in the state Legislature, I felt called to serve my country in the darkest days of the war. So I voluntarily enlisted to serve in Iraq on the front lines as an operations officer in an intelligence unit. The experience...

• Congressional candidate Dan Roberts rejected family's politics; wounded in Vietnam: Editor's note: This is one in a series about the 2nd District congressional candidates. By Richard Halstead Marine Corps 2nd Lt. Dan Roberts was leading his troops back from a patrol through Vietnam's Elephant Valley near Da Nang in 1966 when a mine exploded, killing one soldier and piercing Roberts' left leg with shrapnel. "I guess there was an element of shock and disbelief. I had these fragments of shrapnel going through my left calf," Roberts said. After his radio...

3. Money: Candidate's fundraising efforts or financial disclosure. "Money" articles must involve indicators of financial support or money. For instance, if an article is about an interest group endorsing a candidate but it does not mention financial support, the article is categorized under Platform. Similarly, if an article is primarily about a candidate discussing policy issues during a fundraiser and does not mention how much money the candidate raised, the article is categorized under Platform.

• DUNCAN WAR CHEST NOW AT $200,000: State Sen. Jim Duncan has made a strong financial start in his campaign to unseat Republican incumbent U.S. Rep. Don Young in November, reports filed with the Federal Election Commission show. Duncan has raised more than $200,000, with four months to go until his first election test, the August Democratic primary. Money counts in a statewide race, and at his current fund-raising pace, Duncan could become the strongest challenger Young has faced in years.Young has collected three...

• U.S. House candidates Daines, Smith, Rankin report more than $1M in assets: Diane Benson, Democratic candidate for Congress and mother of an Iraq war veteran who lost his legs there, said Friday that the decision to extend the tour of the Alaska-based 172nd Stryker Brigade is wrong. Benson, in a prepared statement, said U.S. Rep. Don Young should be speaking out about it. "I call on Rep. Don Young to immediately stand up for our Alaskan sons and daughters and demand that our Alaskan families are reunited as planned. We cannot allow our families to suffer...

• Texas-based Super PAC Campaign for Primary Accountability targets US Rep. Spencer Bachus, backs challenger Scott Beason: WASHINGTON – A Texas-based political action committee with $1.6 million cash on hand will be spending some of that money to help defeat U.S. Rep. Spencer Bachus, a 10-term veteran who PAC organizers have targeted because of his longevity and ethics investigation. "Incumbents like Mr. Bachus ... are longtime passengers on the inside-the-beltway gravy train," said Curtis Ellis, a spokesman for the Campaign for Primary Accountability. The entrance of a Super PAC – which can spend...

- Bellavia turns financial disclosure into prodding of Collins: which, in Bellavia's case, is minimal. Bellavia this week released the personal financial disclosure statement he is required to file for his candidacy for the Republican nomination for Congress in New York's 27th district. And the document shows the Iraq War hero with family income so far this year of no more than $11,820, along with a credit card debt somewhere between $15,001 and $50,000.

4. Platform: Information on candidates' policy stance or ideological platform.

- BENTON'S ABORTION VIEWS CHANGE: As he wades ashore in the battle to capture his party's congressional nomination, state Sen. Don Benton has made a major shift in his position on abortion rights. Like the three other Republican soldiers battling for this beach, Benton is now in the pro-life or anti-choice camp. He had been the only pro-choice candidate among the four GOP contenders for the office. The others, Pat Fiske, Paul Phillips and Rick Jackson, were already dug in as opponents of general legalized abortion in...
- Harris wants state gas tax to be suspended; Kratovil calls suggestion irresponsible: The Republican candidate for Maryland's 1st congressional district wants the state gas tax suspended for three months. His Democratic opponent said suspending the gas tax would be irresponsible without coming up with a plan to offset the loss of tax revenues to the state. State Sen. Dr. Andy Harris, R-7th, Baltimore and Harford counties, said Gov. Martin O'Malley should call a one-day special session of the legislature so lawmakers can suspend the state tax on gasoline and diesel...

5. Scandal: Any case in which a candidate's ethics is called into question; investigation; lawsuit; legal dispute; allegation.

- Griffith files suit against campaign manager: Defeated candidate claims some funds not accounted for Parker Griffith, who was defeated in the March 13 Republican primary in his bid to return to Congress, filed a lawsuit Thursday against his former campaign manager, alleging breach of contract and failure to properly account for campaign funds. The suit filed in Madison County Circuit Court contends Griffith hired Huntsville resident Barbara Nash on Jan. 12 to work as his campaign manager in the 5th Congressional District race...
- Paton accuses primary foe of fraud, seeks removal from congressional ballot in Arizona: As the Democratic primary to nominate the replacement for Congressman Mike Ross unfolds, it appears that none of the three Democratic candidates comes close to filling his large shoes. The fundraising reports reveal that Hot Springs attorney Q. Byrum Hurst has the most backing of those willing to donate money. But the more we learn about his background, the more you have to wonder why. Hurst reported raising over $100,000 in the first month of his campaign. His opponents...
- Columbia cops arrest state representative for DUI, weapons possession, Ted Vick is a candidate in race for a new congressional seat: Columbia police officers arrested a state

representative Thursday for driving under the influence of alcohol and the unlawful carrying of a pistol after he was stopped for speeding. S.C. Rep. Ted Vick, D-Chesterfield, was released from the Alvin S. Glenn Detention Center on personal recognizance bonds for the charges. He also was given a ticket for speeding. Vick, 39, is one of several candidates seeking the Democratic nomination for South Carolina's new 7th congressional seat....

6. **Endorsement**: Any article in which the word stem "endors" is found.

Appendix B: Performance of Various Supervised Learning Algorithms

Table A.1 – Precision, Recall, and F-scores for each classification algorithm.

Algorithm	Precision	Recall	F-score
SVM	0.798	0.696	0.734
SLDA	0.780	0.682	0.722
Maximum Entropy	0.702	0.702	0.698
Boosting	0.714	0.626	0.658
Random Forest	0.894	0.592	0.656
Bagging	0.784	0.558	0.602
Neural Network	0.552	0.514	0.520
Tree	0.474	0.486	0.472