Can Fact-checking Prevent Politicians from Lying?

Chloe Lim

May 17, 2018

Abstract

Journalists now regularly trumpet fact-checking as an important tool to hold politicians accountable for their public statements, but fact checking's effect has 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 false causes a 9.5 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 affect candidate behavior.

1 Introduction

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] Yet, despite the popularity, fact-checking has received little scholarly attention: its effects have only been assessed anecdotally and in experiments on politicians holding lower-level offices.

Using a rigorous research design, this paper estimates the effects of fact-checking on presidential candidates for the 2012 and 2016 US Presidential Election. As a baseline design, I use an interrupted time series design to show that a fact checking agency deeming a statement false causes a 9.5 percentage point decrease in the probability the statement is repeated in the future. To eliminate alternative explanations that could confound this estimate, I use a series of robustness checks on the model that weaken the assumption 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 conclusions from the interrupted time series design are not merely the result of politicians failing to repeat statements. I also conduct a placebo test that evaluates the trend in false claims during randomly chosen, hypothetical times 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 for reasons unrelated to the fact-checking itself.

Across research designs, I find a significant effect of fact checking on the probability that politicians repeat false claims. The effects were especially pronounced for candidates in 2016. For Hillary Clinton, the probability of repeating a false-rated

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

claim was reduced by 14.5 percentage point after the fact-check. Similarly, for Donald Trump, the probability of repeating a statement in the future decreased by 9.2 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 et al., 2018), this paper demonstrates an important channel through which fact-checking matters in presidential campaigns. By showing that news organizations affect candidate behavior by scrutinizing and evaluating their public statements, this study contributes to the literature on how news media can hold politicians accountable.

1.1 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 Jr and Strömberg, 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 behavior (Snyder Jr and Strömberg, 2010; Ansolabehere et al., 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 proved 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

[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; O'Sullivan and Geiger, 1995)

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 effects of fact-checking, the question of whether fact-checking can effectively monitor political figures remains unanswered. Few studies have formally assessed the effects 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, Hardy, Winneg, and Jamieson (2013) found that fact-checking has been effective in improving the accuracy of voters' perception of candidates' policy platform. 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 were sent letters before the 2012 election. These letters contained information on the reputation and electoral consequences of receiving a negative rating from a fact-checking outlet. Nyhan and Reifler (2015) found that legislators who had received the letters were less likely to make false-rated claims compared to those who had not received the letter.

My design differs from Nyhan and Reifler (2015)'s approach in two ways: First, while they focus on the effects of fact-checking among state legislators, I analyze the effects of fact-checking among presidential candidates. Second, Nyhan and Reifler (2015) evaluate the effects of informing state legislators of the *possibility* of being fact-checked. In contrast, I examine whether an actual practice of fact-checking affects politicians' tendency to repeat an inaccurate statement.

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. Section 2 describes the data collection procedure in detail. In Section 3, I show that fact-checking plays a significant role in preventing politicians from repeating false-rated claims. To validate these findings, in Section 4, I address two alternative hypotheses according to which changes in candidates' behavior found in Section 3 may be caused by factors that are unrelated to fact-checking, such as "topic switch" or "natural decrease over time". 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 has been successful in deterring presidential candidates from repeating false claims. In Section 5, 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. Section 6 offers several possible explanations for the results. Section 7 concludes.

2 Collection of Fact-checked and Unchecked Statements

I gathered 374 speeches (Clinton: 67, Obama: 105, Romney: 77, Trump: 112) made by presidential candidates for the US Presidential election 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). The speech transcripts were available on C-SPAN and the American Presidency Project at UC Santa Barbara.

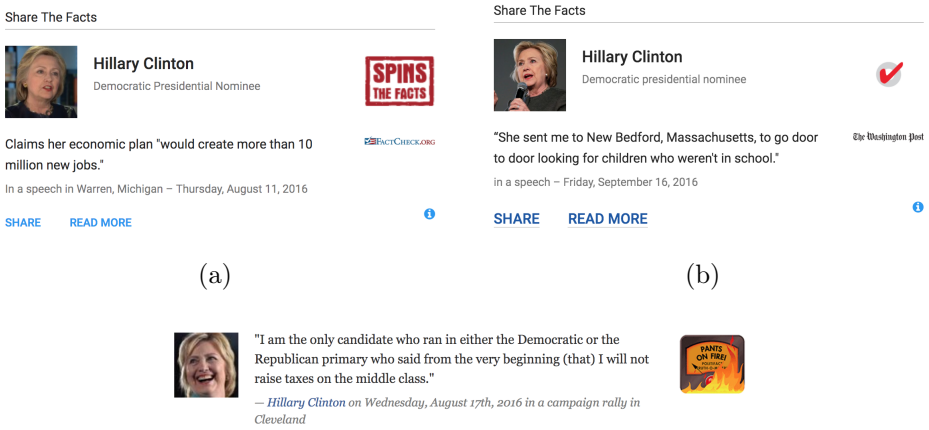
I chose to focus on the period between early August and Election Day to evaluate the immediate effects of fact-checking. As Election Day approaches, speeches are made more frequently, compared to earlier in the year. The average number of days between two consecutive speeches is 6.321 during the primaries[4]. Yet, between August and early November, campaign speeches are made almost daily in rallies and conventions (0.98 speech a day on average). This enables me to observe the immediate effects of fact-checking on candidate behavior in subsequent speeches over a much shorter time span, ensuring that other political events or actions do not 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

[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.

[4]From January to late July of each election year.

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.



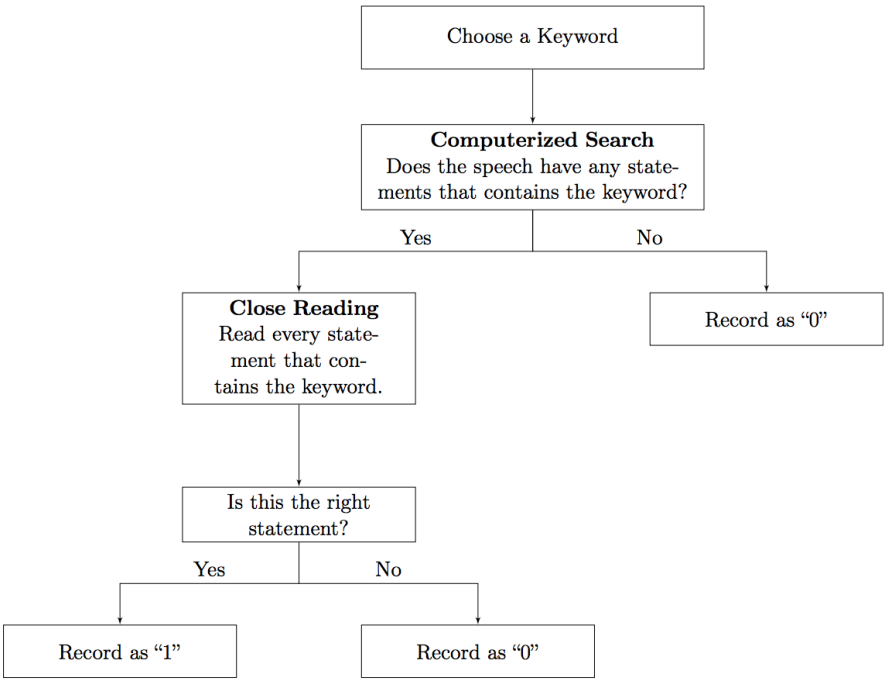
outlets. Likewise, each unchecked statement represents a direct quote of the candidate's statement that has not been fact-checked by a fact-checker. Unchecked statements were collected from the 374 speeches in my sample.

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.

[5]ClaimBuster is a fact-checking platform created by Hassan, Arslan, Li, and Tremayne (2017). The ClaimBuster scores are obtained from a supervised classification model. The model was trained using 20788 sentences from past general election debates, which were labeled by human coders (Hassan et al., 2017).

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



2.1 Dataset of Fact-checked Statements

2.1.1 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 campaign.[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 made in each speech. I carried out the search in two stages – a preliminary comput-

[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.

erized 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.

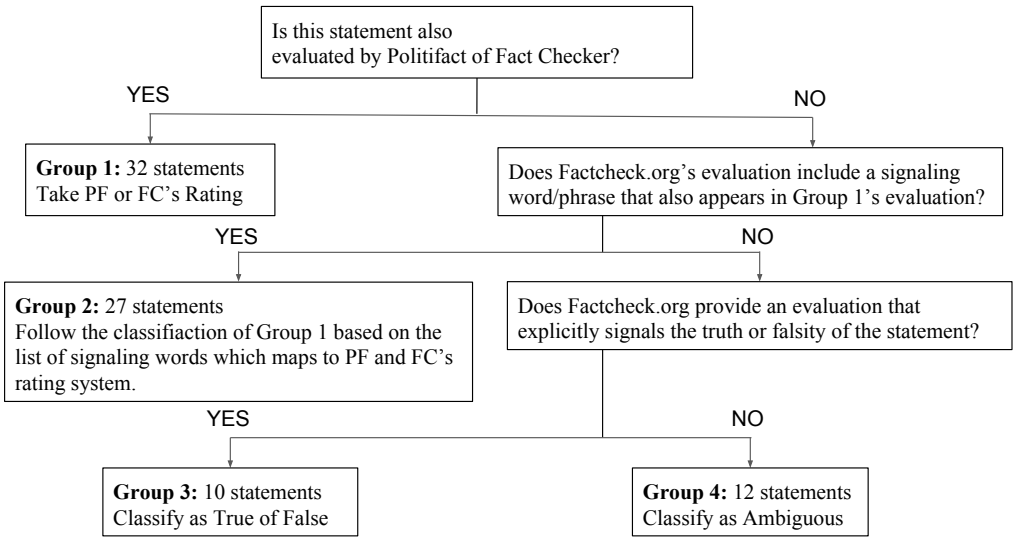
2.1.2 Classification of Ratings

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 (Chief Editor at Fact Checker)'s interpretation of how Fact Checker's scale compares to that of Politifact's, [7], 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 Politifact and "Geppetto Checkmark" and "1 Pinocchio" from Fact Checker. "False"

[7] Kessler writes as follows in his email correspondence with Marietta, Barker, and Bowser (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



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 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-checks (out of 123 fact-checks by 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).

2.2 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 (2016)' 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 Section 4.1, 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).

3 Evaluating the Effect of Fact-checking

3.1 Interrupted Time Series Design for Fact-checked Statements

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 the 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).

Restricting the treatment group to 5 speeches after the fact-check enables me to compute the immediate effects of fact-checking and 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. 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.

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

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

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 $Length_{it}$ – 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. λ_t is a "Time of Speech (relative to the date of fact-check)" fixed effect ($t = 1, \ldots, 10$) that

Table 1 Interrupted Time Series Estimate for False Claims Table 1 displays the interrupted time series estimate for the effects of fact-checking on false-rated claims. On average, fact-checking reduces the probability that a candidate will repeat a false-rated claim by 9.5 percentage points.

	False Claims		
	All	2016	2012
Factcheck	-0.095	-0.101	-0.072
	(0.013)	(0.014)	(0.033)
Statement F.E.	✓	✓	✓
Observations	2,200	1,820	380

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.

3.2 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 false-rated claims on all four candidates, candidates in 2016, and candidates in 2012, respectively. On average, fact-checking reduces the probability that a candidate will repeat a false-rated claim by 9.5 percentage points. Perhaps because fact-checking became more popular and frequent in 2016 compared to the previous election [8], the effect of fact-checking is more pronounced for candidates who ran in 2016, relative to those in 2012. After receiving a "False" from fact-checkers, the probability that candidates in 2016 repeated a given statement decreased by 10.4 percentage on average. For candidates in 2012, although the coefficient on $Factcheck$ is negative, the effect size is smaller, compared to their 2016 counterparts.

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

Figure 4. Percentage of Speeches Featuring False Statements 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 13 percentage points in 2016 and by 9 percentage points in 2012.

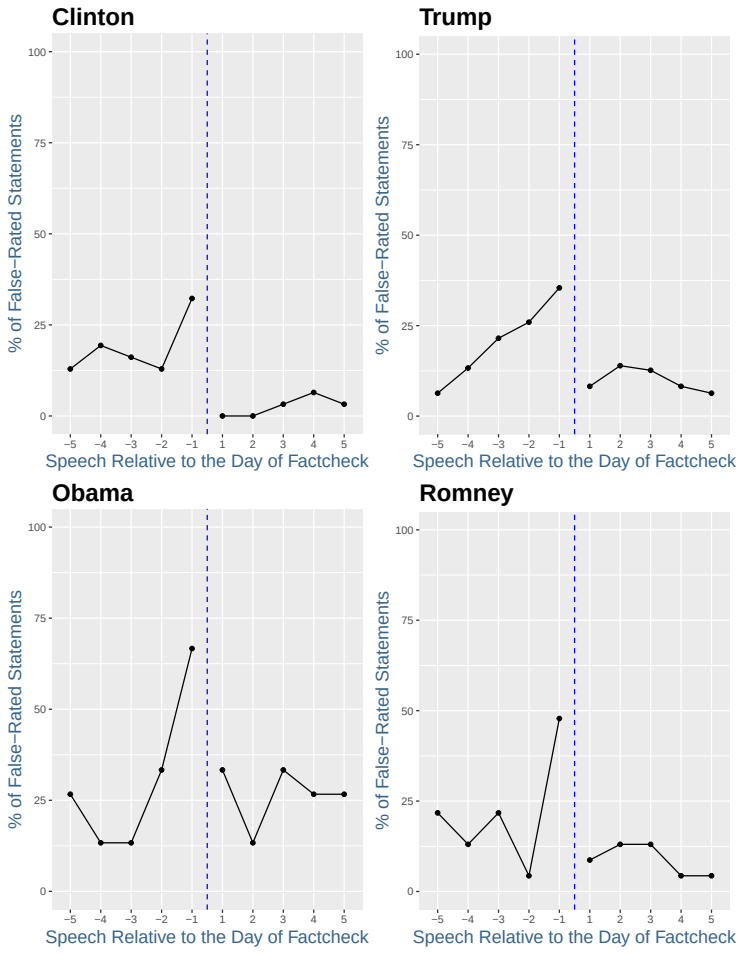


Figure 4 displays the percentage of false-rated statements made in each speech, relative to the day of fact-check. The percentage of false-rated claims per speech is computed as follows: $\frac{\sum_{i=1}^{n} Spoken_i}{n}$ where n represents the total number of false-rated statements for a given candidate 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 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 13.261 percentage points after the fact-check. More specifically, for Clinton, an average percentage of false-rated claims per speech went down from approximately 16 to 1 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 27), compared to before (around 22). Although the effect size varies by year, fact-checking has led to a non-negligible drop in the percentage of speeches that feature a false claim, especially during the 2016 presidential election.

4 Testing Alternative Explanations

In this section, I run additional analyses to eliminate alternative explanations that could confound the results in the previous section. 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.

4.1 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 are: ClaimBuster scores, pre-treatment frequency, and topics.

1. ClaimBuster Scores: As mentioned earlier, ClaimBuster scores indicate "checkability" — a criterion used by fact-checkers when findings facts 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 measures 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 is performed to minimize the difference in pre-treatment frequency.[9]

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 classifies 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 is 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, "Jonathan

[9]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.

Gruber (architect of Obamacare) said the American people are essentially stupid for approving and allowing Obamacare to happen" is assigned to Topic B - Immigration and Foreign Affairs. Under the manual classification, the statement is 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) are assigned to "Topic C - Economy and Campaign" under both LDA 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 is 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 203 matches of checked and unchecked statements. On average, a checked statement is matched with 1.2 unchecked statements. Of the 203 matches, there were 175 cases in which there was no difference in pre-treatment frequency; 25 matches differed by 1 and 3 matches differed by 2. The mean difference in ClaimBuster scores between checked and unchecked statements is 0.003 (p-value: 0.3223). There were 18 cases in which the matched statements had different topic pairings but were still close in terms of ClaimBuster scores and pre-treatment frequency. Only one fact-checked statement was 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.

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

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 played a role in preventing candidates from repeating false-rated claims.

	Unchecked vs. Checked		
	All	2016	2012
Factcheck	-0.098 (0.020)	-0.093 (0.021)	-0.121 (0.052)
Statement F.E.	✓	✓	✓
Time of Speech F.E.	✓	✓	✓
Observations	3,840	3,110	730

The difference-in-differences estimate is negative for both 2012 and 2016, which implies that fact-checking had an effect in reducing 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.

4.2 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 have avoided repeating false-rated statements, not because these statements were found to be false, but because they simply moved onto a new agenda or no longer felt 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 columns 1-3 of Table 3). I then compare how post-factcheck speeches differ from pre-factcheck speeches for "True" versus "False" state-

Table 3 Interrupted Time Series Estimate for True Claims & Diff-in-Diff Estimate for True vs. False Claims Table 3 displays the interrupted time series estimate for the effects of fact-checking on true-rated claims (columns 1-3) and difference-in-differences estimates for true vs. false claims (columns 4-6). 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 Claims			True vs. False		
	All	2016	2012	All	2016	2012
Factcheck	-0.064	−0.080	−0.048	−0.026	−0.006	−0.038
	(0.026)	(0.033)	(0.041)	(0.027)	(0.036)	(0.052)
Statement F.E.	✓	✓	✓	✓	✓	✓
Time of Speech F.E.				✓	✓	✓
Observations	660	360	300	2,610	1,930	680

ments.[10] The p-value for the difference in pre-treatment frequency between "True" and "False" claims is 0.588. This is evidence in favor of the parallel trends assumption.

As shown in columns 1-3 of Table 3, although the magnitude on the coefficient for true-rated claims is slightly 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 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".

Columns 4-6 of Table 3 show difference-in-differences estimates for true-rated vs. false-rated claims with statement fixed effects for all 4 candidates and candidates

[10] "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.

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, especially for candidates in 2016. Yet, the sign of the difference-in-differences coefficients are negative for all candidates. This implies that a negative score from fact-checkers did play a role in preventing candidates from repeating false-rated claims, at least in some cases.

4.3 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.

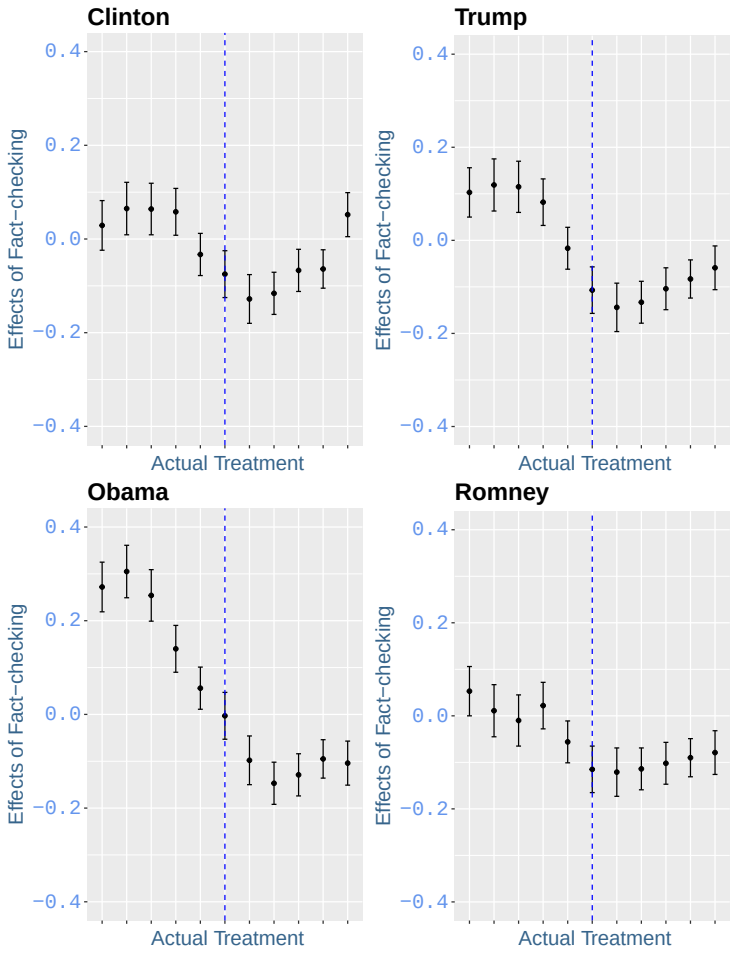
For each round of analysis, a corresponding "placebo" date is used to assign speeches into a treatment and control group. Because there are 10 such "placebo" dates for each fact-check, the following fixed effects regression is run 10 times:

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

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.

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 Clinton, Trump, and Romney, 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.



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, 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, re-

spectively, as a placebo treatment date (See Appendix E for estimates of coefficients and standard errors for each placebo date).

As shown in Figure 5, it seems that Obama does not respond as quickly to fact-checks since the sign of the coefficients remain positive until the day after the actual date of fact-check is used as a placebo date. For Clinton, Trump, and Romney, 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. For these three candidates, the magnitude of the negative coefficient on an actual treatment 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.

4.4 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

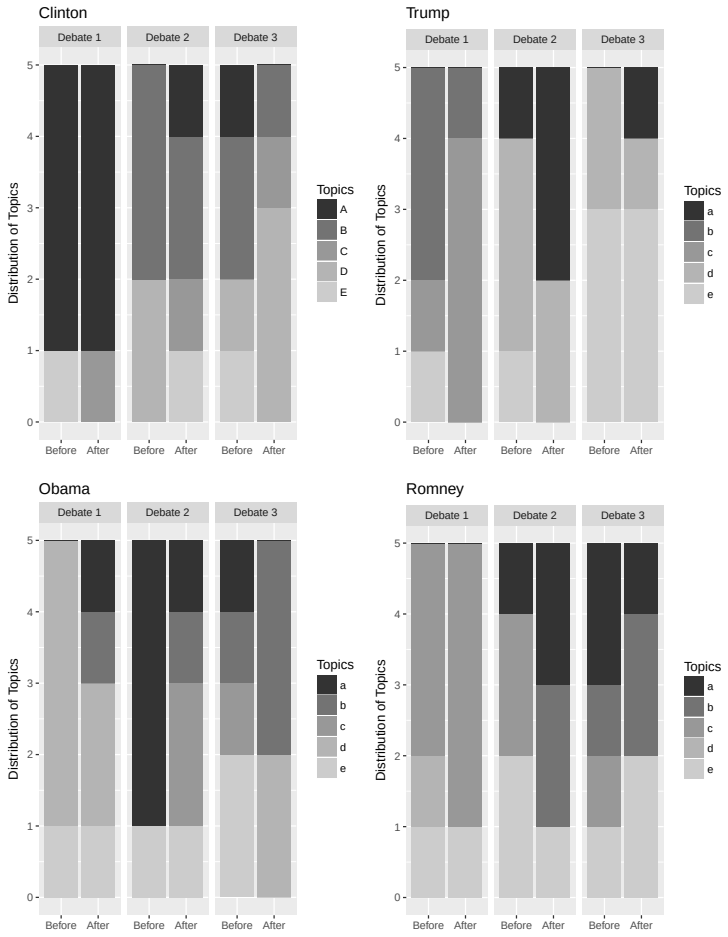
the fact-checking itself.

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 when 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

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.



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.

5 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 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, whose average number of views was around 3,000. Presidential debates are also the busiest time of year for fact-checkers. 41.5 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 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 4 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

Table 4 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	-0.041	−0.019	−0.180	0.005	0.030	−0.076
	(0.026)	(0.028)	(0.041)	(0.036)	(0.041)	(0.081)
Statement F.E.	✓	✓	✓	✓	✓	✓
Speech F.E.	✓	✓	✓	✓	✓	✓
Observations	228	191	37	228	191	37

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 4 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.

6 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 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 caused a 9.5 percentage point decrease in the probability 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 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,[11] 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 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, fact-checking was quite successful in preventing candidates from repeating claims that have been found to be false. Why does fact-checking seem to affect 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 of 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.

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

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.[12]

Third, another possible explanation for the effects of fact-checking may be that sometimes, candidates might 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, presidential candidates do respond to negative fact-checks by reducing the number of false-rated claims in speeches made after the fact-check. The results imply that fact-checking has a much bigger 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). By successfully deterring presidential candidates from repeating "false" claims, fact-checkers have taken steps in the right direction to fulfilling the democratic ideal of political watchdog. In showing that news organizations affect candidate behavior, this study contributes to the literature on how news media can hold politicians accountable.

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

References

Ansolabehere, S., E. C. Snowberg, and J. M. Snyder (2006). Television and the incumbency advantage in us elections. *Legislative Studies Quarterly 31*(4), 469–490.

Barrett, P. (2001). Interrater reliability: Definitions, formulae, and worked examples.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

Cappella, J. N. and K. H. Jamieson (1994). Broadcast adwatch effects: A field experiment. *Communication Research 21*(3), 342–365.

Carr, D. (2012). A last fact check: It didn't work.

Gottfried, J. A., B. W. Hardy, K. M. Winneg, and K. H. Jamieson (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.

Guess, A., B. Nyhan, and J. Reifler (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *The Knight Foundation*.

Hassan, N., F. Arslan, C. Li, and M. Tremayne (2017). Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1803–1812.

Lim, C. (2018). Checking how fact-checkers check.

Marietta, M., D. C. Barker, and T. Bowser (2015). Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? In *The Forum*, Volume 13, pp. 577–596. De Gruyter.

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

Nyhan, B. and J. 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, P. B. and S. Geiger (1995). Does the watchdog bite? newspaper ad watch articles and political attack ads. *Journalism & Mass Communication Quarterly 72*(4), 771–785.

Snyder Jr, J. M. and D. Strömberg (2010). Press coverage and political accountability. *Journal of political Economy 118*(2), 355–408.

Stencel, M. (2015). Politicians modify words, prepare evidence to satisfy fact-checkers.

Wintersieck, A. L. (2017). Debating the truth: The impact of fact-checking during electoral debates. *American Politics Research 45*(2), 304–331.