A Dynamic Measure of Bureaucratic Reputation: New Data for New Theory

Luca Bellodi*

Abstract

Bureaucratic reputation is one of the most important concepts used to understand the behaviour of administrative agencies and their interactions with multiple audiences. Despite a rich theoretical literature discussing reputation, we do not have a comparable measure across agencies, between countries, and over time. I present a new strategy to measure bureaucratic reputation from legislative speeches with word-embedding techniques. I introduce an original dataset on the reputation of 465 bureaucratic bodies over a period of forty years, and across two countries, the US and the UK. I perform several validation tests and present an application of this method to investigate whether partisanship and agency politicisation matter for reputation. I find that agencies enjoy a better reputation among the members of the party in government, with partisan differences less pronounced for independent bodies. I finally discuss how this measurement strategy can contribute to classical and new questions about political-administrative interactions.

Forthcoming, *American Journal of Political Science*

Keywords: Bureaucracy, Bureaucratic Reputation, Natural Language Processing

Word count: 9,953

*PhD Candidate,UCL Political Science and Research Fellow, DONDENA Center, Bocconi. For helpful comments, I thank Tom O'Grady, Colin Provost, David Coen, Arthur Spirling, Pedro Rodríguez, Marco Schito, Chase Foster, and participants at the Yale American Politics Graduate Workshop. I also wish to thank three anonymous reviewers and the editor for insightful feedback and Jack Blumenau for sharing a development version of the UK House of Commons speech data.

Introduction

The political science literature has made important strides to enhance our understanding of the drivers of bureaucratic behaviour and the sources of bureaucratic power. Ever since the first models of political control of the bureaucracy, scholars have mostly focused on structural features of the bureaucracy, namely formal discretion, administrative procedures, and oversight mechanisms, considered to be the main tools to control bureaucratic policy-making (Epstein and O'Halloran 1999; Huber and Shipan 2002; McCubbins, Noll, and Weingast 1987; Moe 1990). Yet later advancements in the scholarship have gradually crumbled this structuralist approach to the study of the bureaucracy, and interpreted bureaucratic behaviour as unfolding through formal and informal channels or, in other words, implicit and explicit contracts (D. P. Carpenter and Krause 2015). One of the most prominent attempts to theorise what happens beyond formal contracts is Carpenter's reputation-based account of bureaucratic autonomy (D. P. Carpenter 2001, 2010).

It is thanks to bureaucratic reputation – "a set of symbolic beliefs about an organisation embedded in a network of multiple audiences" – that agencies become autonomous actors and manage to secure the policies they favour despite the opposition of even the most powerful politicians (D. P. Carpenter 2001, 3–4). During the US progressive era, well-esteemed bureaucracies such as the US Post Office Department and the Department of Agriculture were indeed consistently able to induce Congress and the President to consider and pass legislation that was quite different from their original preferences. This and subsequent works on bureaucratic reputation inaugurated a new tradition of scholarship that integrates formal (e.g., structure, capacity, and procedures) and informal accounts to better understand the interactions between the bureaucracy and other political actors. However, measuring reputation is a daunting task and scholarly work has been limited to a one-country, one-agency, one-year approach, failing to match the innovative scope of bureaucratic reputation as a new prominent account of bureaucratic politics.

For bureaucratic reputation to become a general theory of bureaucratic behaviour and to talk to other subfields in political science, we need empirical work that is able to identify the effects and causes of reputation more systematically. If we want reputation to talk to theories of delegation (Thomson and Torenvlied 2011), interest groups (Nelson and Yackee 2012), political oversight (Lowande 2018), and rule-making (Potter 2017), we need measures that allow theories of reputation to be jointly tested with and against alternative explanations of administrative outcomes. In this paper I build on recent advancements in natural language processing and propose a new method to measure bureaucratic reputation from millions of speeches given by politicians in parliament. I employ word-embedding techniques to understand how politicians talk about bureaucratic agencies and derive word vector representations for each agency in every year and then measure the distance of these vectors from a vector that captures positivity. Agencies with better reputations will be "closer" to this positivity embedding than agencies with worse reputation. I introduce an original dataset on bureaucratic reputation for 465 agencies across two countries – the US and the UK – and over almost forty years. I use both quantitative and qualitative information to demonstrate the validity of these measures, showing that the estimates react meaningfully to important changes or scandals that involved the agency, and that they positively correlate with related measures such as public opinion about government agencies.

I present an application of this method to politicians' polarisation when they talk about bureaucracy. By splitting the initial corpus of speeches by political party, I estimate partisan measures of reputation and show how reputation differs by political party, party in government, and agency politicisation. I show that agencies enjoy a better reputation among the members of the party in government, but the difference between reputation among majority and opposition party members is smaller for independent agencies and non-departmental bodies. Finally, I discuss some limitations of the proposed measurement strategy and outline directions for future research.

This measurement strategy opens new paths to the study of political-administrative interactions, offering new data for both classical questions on control and delegation, as well as new questions that integrate theories of bureaucratic politics with insights from other subfields of political science such as partisan identity and political polarisation.

The Need for a New Measure of Reputation

Bureaucratic reputation has informed an innovative literature that bridges the principal-agent paradigm with informal political-administrative dynamics. Krause and Douglas (2005), for instance, show that presidential, congressional, and independent regulatory commissions are concerned with homogenous reputational considerations about the quality of their decisions that outweigh the different political pressures resulting from their degree of insulation from other political actors. Maor (2007) describes agency independence and its scientific gold standard as reputation protection mechanisms, able to legitimise bureaucratic decisions, and Krause and Corder (2007) find that bureaucracies under tight political control make less accurate (and more optimistic) economic forecasts because they discount future reputation costs associated with their mistakes at a steeper rate than independent organisations.

Works on bureaucratic accountability have employed reputation-based accounts too. E. M. Busuioc and Lodge (2016), for instance, conceive accountability as the practice of sustaining and cultivating reputation across multiple audiences, far beyond formal requirements aiming to reduce informational asymmetries and agency slack. Empirical support for this theoretical claim is offered by Gilad, Maor, and Bloom (2013), who showcase how reputation explains the communication strategies of agencies facing media attacks: silence in domains where the agency is well esteemed, and attention and responsiveness where reputation is weak. Reputation ultimately allows agencies to "generate public support, to achieve delegated autonomy and discretion from politicians, to protect the agency from political attack, and to recruit and retain valued employees" (D. P. Carpenter 2002, 491).

However, while bureaucratic reputation has been put forward as a new currency of bureaucratic politics, able to explain autonomous policy-making, delegation, and accountability, the discipline still lacks a comparable measure of reputation across agencies, between countries, and over time. While the number of empirical works on reputation has increased in the last few years, scholars still employ either qualitative data (E. M. Busuioc 2016; Gilad and Yogev 2012) or quantitative proxies such as the valence of media coverage (Maor and Sulitzeanu-Kenan 2013). When measured quantitatively, reputation has mostly been studied from the supply-side, namely by looking at how agencies respond to reputational threats (Maor, Gilad, and Bloom 2013) or try to manage their reputation through external communication. Lee and Whitford (2013), for instance, measure reputation as the number of freedom-of-information-act request denials and the time to respond to a request, L. J. Anastasopoulos and Whitford (2019) look at twitter profiles of agencies, and M. Busuioc and Rimkutė (2019) perform a quantitative content analysis of agency annual reports.

However, reputation is a set of beliefs among audiences and it therefore seems appropriate to measure bureaucratic reputation from the demand-side, trying to understand what these audiences' beliefs look like rather than what the agency does to change them. A recent measurement strategy that addresses this issue and makes use of survey instruments has been devised by Wood, Overman, and Busuioc (2020), who produce a detailed and multidimensional estimate of agency reputation based on a systematic sampling of key stakeholders. Yet the results are still limited to one agency – the EU Chemicals Agency – at one specific time, and in a single political system – the European Union. Furthermore, surveys about technical issues such as bureaucratic agencies tend to induce answers (Zaller 1992). When individuals are asked what they think about the Pensions Regulator, they might not even know that it exists.

Clearly, there is a trade-off between nuance and multidimensionality, on the one hand, and time-, country-, and agencies-coverage, on the other hand. In this paper I address this gap by proposing a new quantitative measure of reputation that trades nuance with coverage, while at the same time focusing on a key audience, politicians, that represent a good synthesis of what a diverse set of audiences perceive the bureaucracy to be. The main assumption invoked is that politicians' electoral incentives to align with key stakeholders when engaging with bureaucratic agencies make them a good - though partial - source of information to capture agencies' reputation, a source that includes politicians' voice too – one that has been surprisingly neglected.

The Whereabouts of Reputation: Audiences and Beliefs

The standard definition of bureaucratic reputation in political science is Carpenter's (2010, 45), for which reputation is "a set of symbolic beliefs about the unique or separable capacities, roles, and obligations of an organization, where these beliefs are embedded in audience networks." The key words of this definition are *beliefs* and *audience networks*.

Every attempt at measuring reputation will therefore have to deal with two questions: what are the beliefs that convey information about reputation?, and what are the audiences whose beliefs give shape to bureaucratic reputation?

The first question is trivial. Beliefs are the value judgements or perceptions about various traits of an organisation. The literature generally clusters beliefs into four main facets or reputation, performative, moral, procedural, and technical (D. P. Carpenter 2010), but I follow the general conceptualisation of empirical work and interpret reputation as an aggregate unidimensional measure of the perceptions of multiple audiences about an agency that spans between a positive and negative extreme.

The question "what audiences should we care about?" conversely, is a hard one. Ideally, measures of reputation would start from an accurate mapping of the various constituencies that qualify as audiences, and would then measure how positive or negative each audience's beliefs about the agency are. This exercise – the one generally used in surveys – besides being very costly, is also inherently arbitrary, for a decision to consider farmers' associations an audience of the Department of Agriculture would itself end up being a stand-alone empirical question. This is why scholars have incorporated the identification of the agency's key audiences in the research question itself, using historical analysis, elite interviews, or secondary sources to map an agency's audiences, their beliefs, and various policy outcomes (E. M. Busuioc 2016; D. P. Carpenter 2001, 2010; Gilad and Yogev 2012; Maor and Wæraas 2015). An alternative solution to the audience identification problem advanced in the literature is content analysis of the news. Yet newspapers may not be the most appropriate venue to look for perceptions and beliefs about administrative bodies. In addition, media coverage is mediated by editorial concerns and likely to suffer from selection bias, for agencies are more likely to end up on the news when problems, scandals, or clear inefficiencies afflict their related sector.[1]

The challenge is then to devise a measurement strategy that retains the attention to the multiplicity of audiences of qualitative works, that continues the attempt of time-varying measurements of early quantitative works, and that also allows for cross-country, -policy, and -agency comparisons. We need an alternative venue where the perceptions of multiple audiences are voiced and can be condensed into a unidimensional measure without deciding which audiences are part of the sample. I propose that legislatures are close to this ideal venue. Legislatures are the right place to capture what audiences – as mediated by their representatives – think of an agency while at the same time letting politicians decide which these audiences are.

Reputation exists at audience-level but only key organised interests qualify as legitimate audiences. Politicians have an interest – because of genuine policy/ideological motivation, strategic re-election interest, or both – to represent these groups and voice their beliefs about the agency. What politicians say during legislative debates can then be a good source of information to measure reputation. It seems plausible to think of politicians as the messengers of external audiences such as business actors, trade unions, consumers associations, or non-for-profit organisations. Free-market parties can report the complaints of businesses towards the alleged over-regulation of the Environmental Protection Agency, whereas social-democratic parties can lament the loose regulation of the Financial Services Authority.[2] What politicians say is then a good proxy not only of what the audiences say, but what politically salient audiences say. It is rational for politicians to align or to report what salient groups think about an organisation. Legislative speeches might even be seen

[1]Parliamentary debates too can suffer from selection bias, but the fact that politicians have to regularly discuss a wide range of policy issues results in more uniform coverage.

[2]Barberá et al. (2019), for instance, find that US legislators are more likely to follow, than to lead, discussion of public issues.

as a device to sort audiences' perceptions by political salience. Irrelevant audiences will be less likely to be dedicated attention by politicians who seek the support of the electorate.

Clearly, legislators are not neutral messengers able to precisely map all the relevant audiences of an agency. They might give voice disproportionately to some audiences or mischaracterise some audiences' beliefs. However, the diversity of their professional and personal background, as well as the interests and geographical areas they represent make legislators a rich and inclusive source of information about an agency's reputation, which minimises researchers' discretion and ensures multiple audiences are voiced.

Parliamentary speeches are frequently used as data input for empirical constructs. Beyond ideological scaling, speeches have been used to estimate the political agenda (Quinn et al. 2010), political influence of MPs (Blumenau 2019), and speeches' complexity (Spirling 2016), to name but a few. Tapping into what politicians say can thus be used in bureaucratic politics too to understand how politicians and elites talk about agencies and therefore estimate bureaucratic attributes such as reputation.

Agencies as Word Embeddings

Recent advancements in machine learning and natural language processing allow researchers to devise new measurement tools to study the bureaucracy (on delegation, see e.g. J. Anastasopoulos and Bertelli 2019; Vannoni, Ash, and Morelli 2020). In particular, a new set of techniques called "word embeddings" – first developed in computational linguistics to learn about semantics (Pennington, Socher, and Manning 2014) – offer new frontiers to measure bureaucratic attributes from text data. The core idea at the basis of word embeddings is that we can "know a word by the company it keeps" (Firth 1957, 11), and we can therefore derive its meaning from the context in which the word is used.

Word embeddings are technically the coefficients from neural network models that predict the occurrence of a word by the surrounding words in a textual sequence. A word of interest is represented as a dense, real-valued vector of numbers, whose length is informative about the complexity of the multidimensional space in which the word is embedded, and whose elements convey information about the semantic meaning of the word, with distances between such vectors capturing how similar the words are (Spirling and Rodriguez 2019). For instance, if the distance between the vector representation of the words "market" and "inequality" is smaller for social-democratic parties than for conservative parties, we can learn the views of market economy of different party families. Similarly, by looking at the vector representation of words that are most similar to the vector representation of word "women," we can examine how individuals, groups, or parties think about the role of women in society. The key innovation of word embeddings is that the meaning of the words is something that is learned from the text and is not exogenously given as in other text analysis approaches that look at word frequency.

Word embeddings have recently entered published work in political science. Preotiuc-Pietro et al. (2017), for instance, use word embeddings to estimate ideology based on tweets and to identify politically moderate and neutral users, and Rheault and Cochrane (2019) fit models of word embeddings augmented with political metadata to estimate the ideology of parties and politicians. The flexibility of word embeddings has been also used to address some limitations of sentiment analysis (Rice and Zorn 2019), and to study how ethnic and gender stereotypes change over time (Garg et al. 2018). The novel measurement strategy I propose builds on these recent trends and represents the first attempt to use large parliamentary corpora and word embeddings in the study of bureaucratic politics.

Countries, Speeches, and Agencies

The textual corpus from which I estimate word embeddings are all the legislative speeches from 1980 to the most recent available data in the two chambers of the US Congress and the UK House of Commons.[3] I decided to focus on the US and the UK because while they both have highly competent bureaucracies in charge of the implementation and administration of policies, the UK civil service is deemed to be neutral and merit-based, while US agencies are on average more politicised (Hood 1991; Lewis 2008). This makes the study of reputation in these two cases informative about the relationship between bureaucratic traditions and reputation. As for more practical reasons, both the UK and US speeches are easily accessible and in the same language, making the estimation procedures less complicated. I set the time-frame from 1980 because the meaning of the words I use in the estimation has not changed since then, and it is therefore possible to compare estimates over time. Finally, a time coverage of about 40 years is a good balance between allowing reputation to change over time while ensuring an accurate mapping of all the agencies and their multiple denominations.

US congressional speeches were downloaded from the Social Science Data Collection of Stanford University (Gentzkow, Shapiro, and Taddy 2018), while UK parliamentary speeches were downloaded from UK Data Service ReShare (Blumenau 2021), for a total of almost 4,9 million speeches (2.52 mln speeches for the US and 2.37 for the UK). I created a list of agencies as comprehensive as possible from both existing datasets and government official websites, for a total of 636 bureaucratic bodies, 285 for the US and 351 for the UK.[4]

[3]I exclude the speeches from the House of Lords because the House of Lords has different functions from the House of Commons, and Lords are unelected, therefore the assumption for which politicians have an incentive to represent key constituencies does not hold.

[4]For the US, I used the samples in Bertelli et al. (2013) and Selin (2015). For the UK, I created a list of agencies from gov.uk/government/organisations.

GloVe

To estimate word embeddings, I employ the unsupervised learning algorithm GloVe (Pennington, Socher, and Manning 2014), a count-based model that produces vector representations of words by doing dimensionality reduction on a co-occurrence matrix. The first step is to create a term co-occurrence matrix X of dimension $V \times V$, where V is a vocabulary consisting of all the unique tokens that appear in the corpus. Each element X_{ij} is a number representing how many times word i co-occurs in the context of word j, with the context simply being a pre-defined window of words whose size depends on the particular task at hand. For example, if word $i = FED$, word $j =$ *policy*, and the window size is symmetric and equal to six, X_{ij} is the number of times FED (target word) co-occurs within six words to the left and right of the word *policy* (context word). Let X_j be the sum of the co-occurrences of any word i with the context word $j =$ *policy* (i.e., the sum of the j^{th} column), and $P(i|j) = X_{ij}/X_j$ be the probability that word i appears in the context of word j.

While the technical aspects of Glove are complicated, the main idea is not. The intuition is that we can learn about the relationship between words and discriminate between words related to one word but not another by looking at the ratio of co-occurrence probabilities. Suppose we want to learn the relationship between the words FED and EPA in a year when the EPA is highly criticised. To do so, we compare the probabilities of these two words happening with various probe words k. We might expect word $k =$ *independence* to be related to the word FED more than to the word EPA, and word $k =$ *critic* to be related more to the word EPA than to the word FED. Similarly, we expect the word $k =$ *dog* to be related to neither, and the word $k =$ *policy* to be related to both. Table 1 represents these expectations in terms of hypothetical probabilities. The probability ratio for words related to FED is large, whereas for words related to EPA is low. Words related to both or neither have a ratio that approximates 1, because they do not help discriminate between which word is related to which. This is why, compared to the raw probabilities, co-occurrence ratios are better able to encode relevant semantic relations and to understand which words are related to the words FED and EPA.

Probability	$\boldsymbol{k}$ **= independence**	$\boldsymbol{k}$ **= critic**	$\boldsymbol{k}$ **= dog**	$\boldsymbol{k}$ **= policy**
$P(k\|FED)$	.1	.01	.001	.15
$P(k\|EPA)$	.01	.1	.001	.15
$P(k\|FED)/P(k\|EPA)$	$.1/.01 = 10$	$.01/.1 = 0.1$	$.001/.001 = 1$	$.15/.15 = 1$

Table 1: Example of co-occurrence probabilities for target words FED and EPA with related and unrelated context words. Only in the ratio does noise from non-discriminative words like dog and policy cancels out, so that large values correlate well with words associated with FED, and small values correlate well with words associated with EPA.

Word vectors are then estimated with a neural network, namely a statistical model containing one layer of latent variables (the dimensions of the word vectors) between the textual input (term co-occurrence matrix) and the output data (the word vectors). The innovation of Glove compared to other algorithms is that the model is trained on all the non-zero entries of the matrix rather than on the entire sparse matrix or on individual context windows (Pennington, Socher, and Manning 2014). To avoid the model from weighting all the co-occurrences equally, word vectors are estimated for every word in V by training a log-bilinear model with a weighted least-squares objective that tries to predict the context word j in which word i is used. Very summarily, the model minimises the following equation J,

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - log(X_{ij}))^2$$

where $V = \{v_1, v_2, ..., v_V\}$ is the vocabulary, w_i is the vector of the target word, w_j is the vector of the context word, and b_i and b_j are scalar bias terms. $f(X_{ij})$ is a function that determines the weight to each pair of words based on how often they co-occur; pairs of words that co-occur more often will have greater weight. The final output is a word embedding for every word in the vocabulary. For instance, in the 2018 corpus of UK speeches, the five most similar word vectors to the *Home Office* embedding – one of the most mentioned agencies – are *immigration*, *department*, *official*, *windrush*, *minister*, *ask*. From just these six words, we can learn that *immigration* was a key issue for the 2018 post-Brexit Home Office, and that MPs frequently *asked* the *minister* about the *Windrush* scandal.

Application

I train the GloVe algorithm on a local corpus of parliamentary speeches for every year and every country. I follow standard practice in text-analysis and I lemmatise the tokens, remove punctuations, digits, capitalisation, stopwords, and all tokens with two characters or fewer to increase the precision of the estimation. Agencies referred to in more than one way (e.g., CIA and Central Intelligence Agency) were replaced in the text with standardised tokens. I then create a vocabulary with all the tokens appearing at least five times in all the corpus, because words appearing very few times do not convey semantic information. I create a term co-occurrence matrix specifying a window size of 12 tokens and estimate 300-dimensional word vectors with a weighting function $X_{max} = 10$. This means that any pair of words for which the co-occurrence count is greater than 10 will receive a weight of 1, whilst the other weights $w_i \in [0, 1)$.[5]

[5] I estimate the model through 100 iterations, with a convergence threshold of 0.001, and a learning rate appropriate to the size of the corpus, equal to 0.1. I use these parameters because they are deemed to be the most appropriate for semantic tasks (Spirling and Rodriguez 2019). Estimation implemented with the *text2vec* R package.

I then exploit the arithmetic properties of vector representations of words and build a vector that combines some unambiguously positive and negative words that will act as benchmark to measure the reputation of the agency. By deducting clearly negative embeddings from the sum of clearly positive embeddings, I obtain a word vector that captures positivity. The specific word vectors I used are:

$$\vec{positivity} = \vec{successful} + \vec{effective} + \vec{great} + \vec{excellent} \\ - \vec{poor} - \vec{negative} - \vec{terrible} - \vec{bad}$$

where the arrows signify the words are vectors. The selection of words followed four criteria: (1) the meaning of the words should be uncontroversial (i.e., positive or negative), (2) stable over time, (3) similar across countries, (4) present in every local corpus of speeches in any given year and country. The precise words I used are similar to the seed words chosen by Rice and Zorn (2019) to set the benchmark for positivity and negativity dictionaries. In the SI (pp. 17-18) I show that the reputation estimates produced with alternative positivity vectors are highly correlated with the estimates derived from this vector. I finally measure the cosine similarity between the word embeddings of each agency and the $\vec{positivity}$ vector. The reputation score will thus be the angular distance between the two embeddings. Formally,

$$\theta_j = \theta_{(\vec{a},\vec{p})} = \frac{\vec{a} \times \vec{p}}{||\vec{a}|| \times ||\vec{p}||} = \frac{\sum_1^n a_i \times p_i}{\sqrt{\sum_1^n a_i^2} \times \sqrt{\sum_1^n p_i^2}}$$

where $\theta_{(\vec{a},\vec{p})}$ is the cosine similarity between the agency vector $\vec{a}$ and the positivity vector $\vec{p}$, namely the ratio between the sum of the products of the i^{th} elements of the two vectors (the nominator) and the product of the square root of the vectors to the power of two (the denominator). For instance, if the embedding $\vec{FED}$ is semantically very similar to $\vec{p}$, it will have a very high reputation, whereas if the $\vec{EPA}$ embedding is semantically distant, it will have a lower reputation. The resulting metric is normalised to take up values between 0 and 1, where greater values signify better reputation.

Uncertainty

Deriving uncertainty measures from neural network models is an area of research still under development (Rheault and Cochrane 2019). Since I cannot estimate uncertainty based on the variance of θ, I produce upper and lower bounds for every agency-year estimate based on the number of mentions of the agency in any given year. The reputation of agencies mentioned 2,000 times per year will be less uncertain compared with that of agencies that barely happen to be mentioned. I therefore model uncertainty as a reciprocal exponential function of the number of mentions, so that agencies with fewer mentions are penalised but in a non-linear fashion. Mathematically, the upper and lower bounds will be given by $\theta_j \pm |1 - \exp \frac{1}{m_j}|$, where m_j is number of mentions of agency j.[6]

This is clearly a mathematical artefact that nonetheless allows me to estimate uncertainty based on the reasonable assumptions for which the more politicians talk about an agency, the more we can learn about its reputation. For instance, since the vocabulary V consists of words appearing at least five times in the corpus, the least mentioned agency will be mentioned at least 5 times, and it will have upper and lower bounds estimates equal to $\theta_j \pm |1 - \exp \frac{1}{5}| = \theta_j \pm 0.221$. Conversely, for agencies mentioned 1,000 times, upper and lower bounds will be equal to $\theta_j \pm 0.001$.

[6]Given $\frac{1}{m_j} \in (0, 1]$ for any $m_j > 0$, and $\exp(x) > 1$ for any $x \in (0, 1]$, then $\exp(\frac{1}{m_j}) > 1$ and $\lim_{\frac{1}{m_j} \to 0} |1 - exp(\frac{1}{m_j})| \approx 0$ (when m_j is large) and $\lim_{\frac{1}{m_j} \to 1} |1 - exp(\frac{1}{m_j})| = |1 - e| \approx 1.7$ (when $m_j \to 0$).

Results

The final dataset consists of reputation estimates for 465 bureaucratic bodies – 217 in the UK and 248 in the US – and over a period of 39 years.[7] Table 2 reports some descriptive statistics for the dataset as a whole and by country. Full lists of agencies are reported in the SI (pp. 2-16). Agencies are mentioned on average 179 times per year and have a reputation of 0.50, with a standard deviation of 0.15. The average reputation in the two countries is about the same, 0.48 in the UK and 0.52 in the US. The Environmental Protection Agency and the Department of Homeland Security are the agencies with the highest average number of mentions in the US (1,514 and 1,582 times per year, respectively), whereas the Treasury and the Home Office are the most mentioned agencies in the UK (1,829 and 944 times per year on average, respectively).

	Total		**UK**		**US**	
Total Agencies	**465**		217		248	
Observations	**7,067**		2,272		4,791	
Time Coverage (Years)	**40**		40		36	
Variable	Mean	SD	Mean	SD	Mean	SD
Reputation	0.50	0.15	0.48	0.16	0.52	0.15
Mentions	179	344	143	302	196	362

Table 2: Descriptive statistics for all the agencies in the dataset and split by country. The table reports the number of agencies for which reputation estimates are produced, the number of observations and time coverage, and the average reputation both for the total dataset and by country.

Figure 1 and Figure 2 show the reputation of 8 of the most mentioned bureaucratic bodies with the largest year coverage in both countries. At first glance, the comparatively high reputation of the military in the US is remarkable, with the Air Force, the Navy, and the Department of Defense being among the public bodies with the highest reputation over the entire time-frame considered. Though significantly fluctuating, the reputation of

[7] 465 and not 636 as the initial sample because not all the agencies included in the initial lists are mentioned in the speeches.

most of these eight agencies seem to be stationary, as suggested by the loess function in the plot. There are nonetheless important exceptions. After the peak in the early 1980s, the reputation of the EPA, for instance, experienced a gradual but constant decrease. Similarly, the reputation of the Department of Homeland Security drops rapidly after its establishment in 2002.

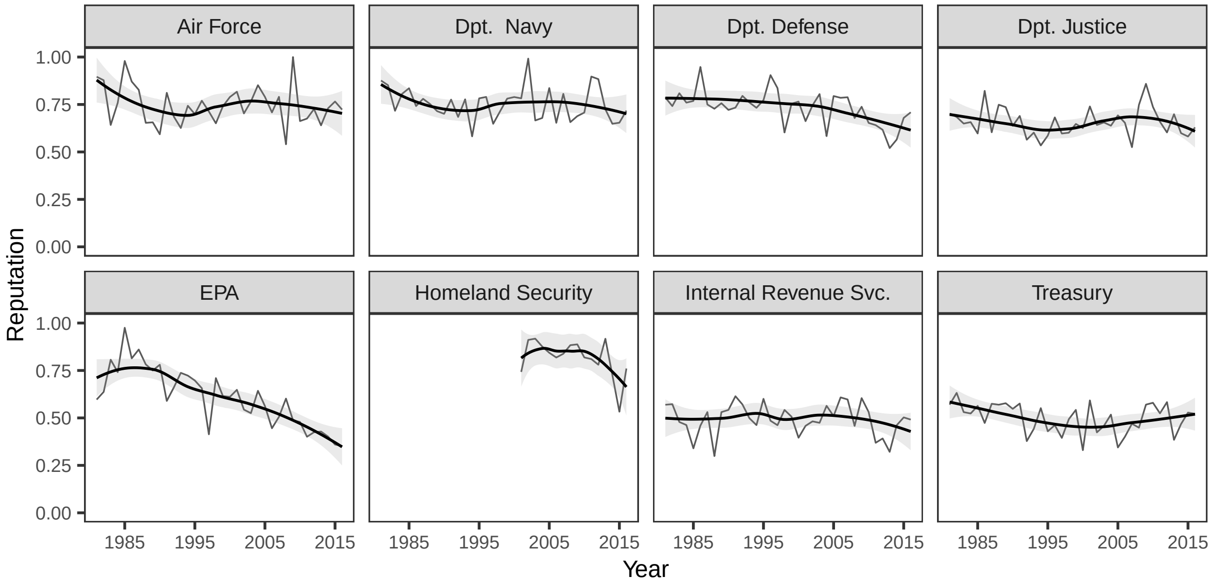

Figure 1: Reputation estimates of US agencies over time with loess approximations superimposed. Estimates are cosine similarity between the "agency" embedding and the positivity vector in every year.

As for the UK, it is possible to see how the reputation of Ministry of Defence drops in 1991, the year of the so-called Options for Change – the dramatic manpower cut to the British Armed Forces after the end of the Cold War – and how the reputation of Network Rail increased in the immediate years after its establishment in 2002 for then rapidly decreasing from 2012 to 2018, possibly as a result of the uninterrupted criticism about delays and service disruptions for commuters.

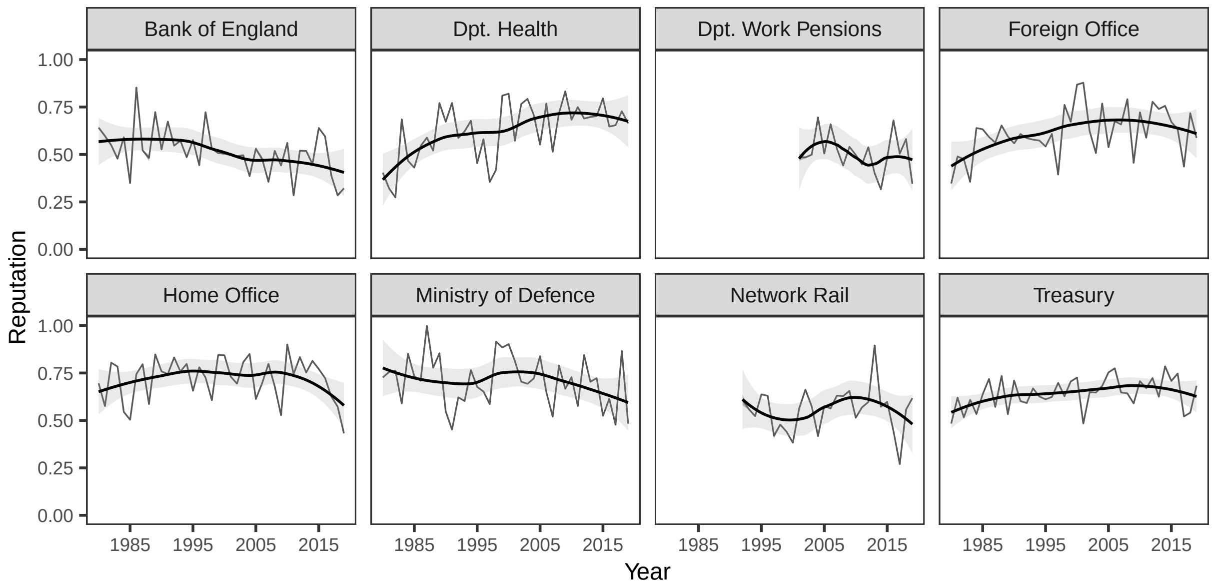

Figure 2: Reputation estimates UK agencies over time with loess approximations superimposed. Estimates are cosine similarity between the "agency" embedding and the positivity vector in every year. Data for Network Rail before its establishment in 2002 are from its predecessor Railtrack, established in 1994.

Figure 3 shows the reputation of the central banks of the two countries, the Bank of England and the Federal Reserve, with dashed vertical lines representing some critical junctures. The FED is overall mentioned more often than the Bank of England, the latter nonetheless enjoying on average a better reputation throughout all the almost 40 years covered by the data (0.52 for the Bank of England and 0.45 for the FED). There is a jump in the reputation of the Bank of England when it became independent, in 1997, while the reputation drops with the 2008 and 2011 financial crises. A similar pattern is followed by the FED. For instance, its reputation falls with financial crises (the one in the early 90s, the sub-prime crisis, as well as the Asian Crisis of 1997), while it increases with the new competences delegated by the Dodd-Frank Act in 2010.

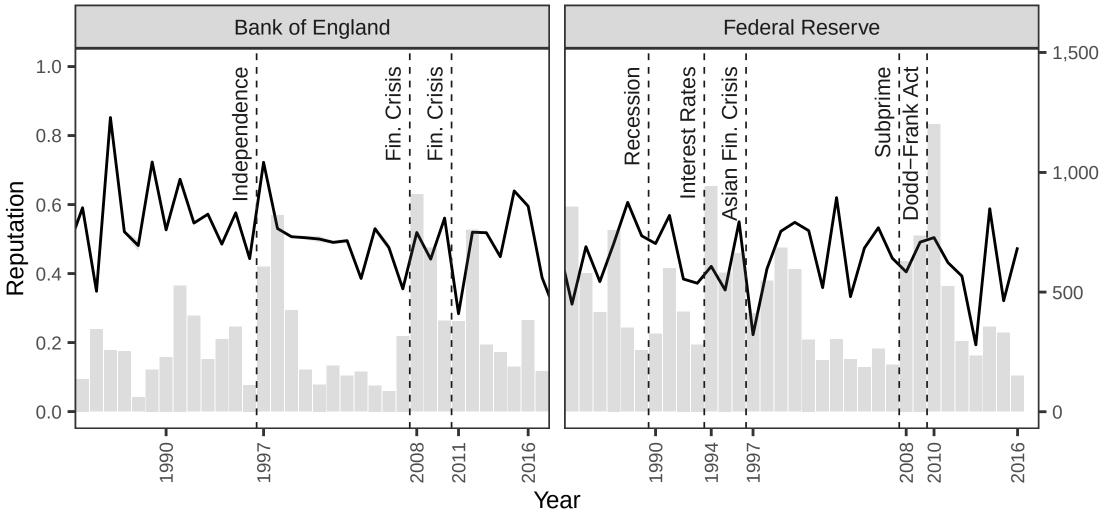

Figure 3: Reputation estimates and number of mentions of central banks over time. Vertical dashed lines represent critical junctures.

Validation

In this section, I address the validity of the measurement with three tests. I assess face and predictive validity with qualitative information about six different agencies to test whether the estimates follow what we would expect the reputation of an agency to be after some critical events. The second test draws from standard convergent validity tests and looks at the relationship between reputation and public opinion (Collier and Adcock 2001). The third test assesses the criterion validity of the measure and compares the reputation estimates derived from legislative speeches with alternative estimates derived from a large corpus of newspaper articles.

When in 2005 Hurricane Katrina hit the state of Louisiana, local and national leaders blamed the poor response of FEMA (Lewis 2008). Yet FEMA's mismanagement was just the second act a of longer play that started in 1992 with Hurricane Andrew, which is "best remembered as an epic bungle by the Federal Emergency Management Agency" (Timeline 2017). Yet in the period between the two hurricanes, the agency had a rapid renaissance. As Roberts (2006, 56) notes, after its reorganisation in 1992, FEMA *"morphed from* *a caricature of the ills of bureaucracy into a model of effective federal administration. Politicians who previously blamed the agency for its slow and inefficient response to disasters came to depend on the agency to lend credibility to their own efforts.”.*

The Department of Homeland Security, after its establishment following the 09/11 terrorist attacks, has been highly criticised over excessive fraud and lack of transparency. Multiple scandals eroded the reputation of the agency. In 2005, the new personnel system called “MaxHR” was blocked in court for violating collective-bargaining employees’ rights (The Washington Post 2008a); in 2008, a Congressional report denounced 15 billion dollars worth of failed contracts (The Washington Post 2008b); and in 2015 the department was found to be operating top secret databases infringing the most basic security procedures (Office of Inspector General (DHS) 2015).

First in 2004, with the dramatic increase in the backlog of pending disability claims, and then in 2014, with the falsified waiting lists, the reputation of the Department of Veterans Affairs has been subject to harsh criticism too (CNN 2014).

Figure 4 plots the reputation of these three agencies over time together with the scandals and critical junctures described above. The reputation estimates capture the decrease in reputation of FEMA after 1992 and 2005, as well as the marked increase after its reorganization in 1992. The reputation of the Department of Homeland Security is consistent with the high criticism, with significant drops and an overall decreasing trend up until 2015, and the reputation of the Department of Veterans Affairs drops in 2004-5 and from 2014 follows a decreasing trend.

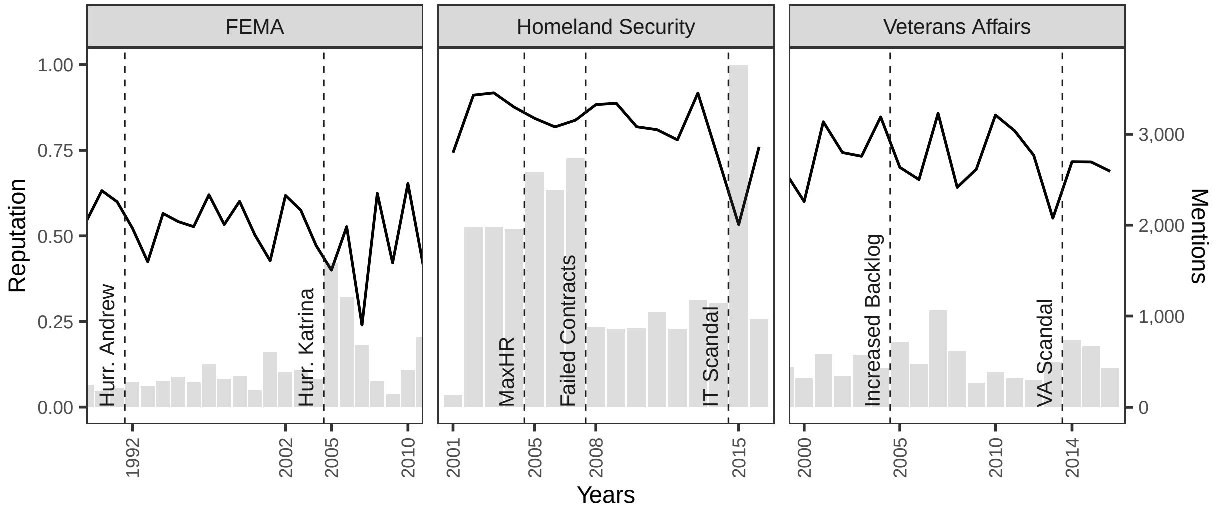

Figure 4: Reputation estimates and number of mentions of US federal agencies and departments. Vertical dashed lines represent scandals or critical junctures.

The reputation of UK agencies too has not been immune to scandals. The Financial Services Authority (FSA) was highly blamed for the 2008 financial crisis, which brought to its abolition later in 2012. The so-called "light-touch" regulatory approach of the authority received cross-partisan criticism (The Daily Telegraph 2008) and was also called into question by the independent review chaired by Lord Turner, which criticised the authority's philosophy for which "markets are in general self-correcting" (FSA 2009, 87). Similarly, the agency's reputation also suffered from the judicial defeat in the *Durant v FSA* case in 2003 – a leading decision with respect to data protection.

Another severe and more recent scandal in British politics involved the Home Office in what has been called the Windrush Scandal. 83 instances were reported in which people were wrongly detained, denied legal rights, and mistakenly deported from the UK (The Times 2018). Although for very different reasons, the reputation of the department was damaged a few years before too, when Jacqui Smith resigned as Home Secretary in 2009 as a consequence of the scandal that involved her husband (The Guardian 2009).

Figure 5 follows quite accurately these events, with the reputation of the FSA reaching one of its lowest levels just after 2003, and dropping again during the 2008 financial crisis and after the abolition announcement of the government in 2010. Similarly, the reputation of the Home Office drops in 2009 and 2018 following the scandals. Finally, the graph also shows a rapid decrease in the Ministry of Defence's reputation from 2009 to 2011, the years of the Strategic Defence and Security Review (Ministry of Defence 2010) – highly criticised by Parliament (House of Commons Defence Committee 2011) – and the Defence Reform Report (Ministry of Defence 2011), which assessed the causes of the department's under-performance and proposed changes to prevent it "from getting into such a poor financial position in the future" (Ministry of Defence 2011, 13).

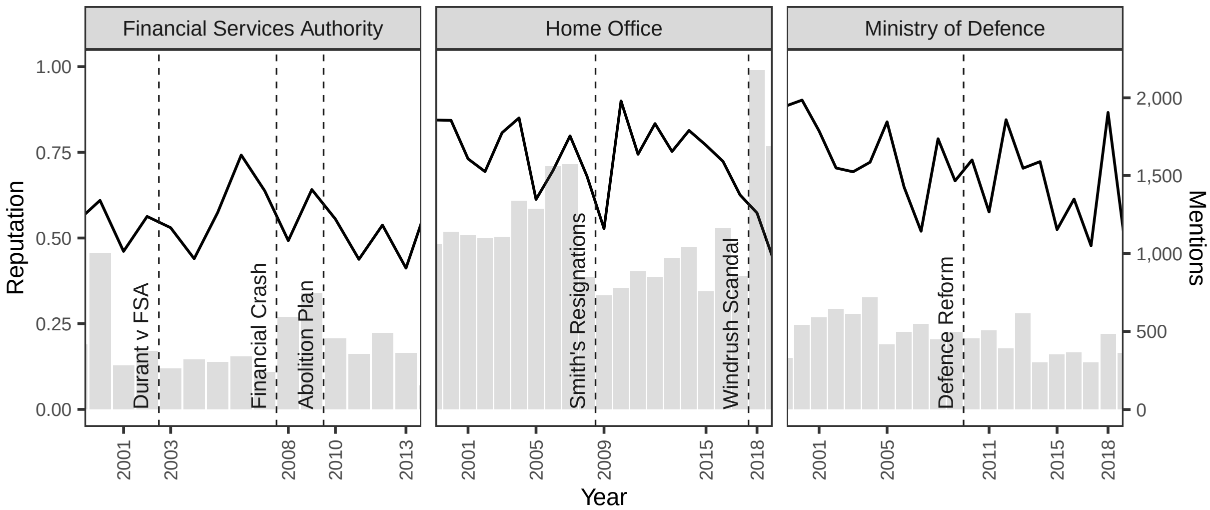

Figure 5: Reputation estimates and number of mentions of UK agencies and departments. Vertical dashed lines represent scandals or critical junctures.

For the second test, I compare the reputation estimates with survey data on public attitudes towards US federal agencies. I assembled a panel of public opinion data from the Pew Research Centre reports (Pew Research Center 2019) and Gallup surveys[8] and matched reputation with public opinion data for 18 agencies over several years, for a total of 51 observations from Gallup and 114 from Pew Research Center. Responses to the survey were recorded on a 4-level scale, from very positive to very negative. Figure 6

[8]Data accessed at the following link: news.gallup.com/poll/27286/government.aspx on 10 February 2020.

plots the relationship between the reputation estimates (x-axis) and the percentage of respondents who reported a very positive or very negative opinion about the bureaucracy (y-axis). Overall there is a positive correlation between reputation and positive opinion about the agency and a negative correlation – though weak – between reputation and negative opinions.

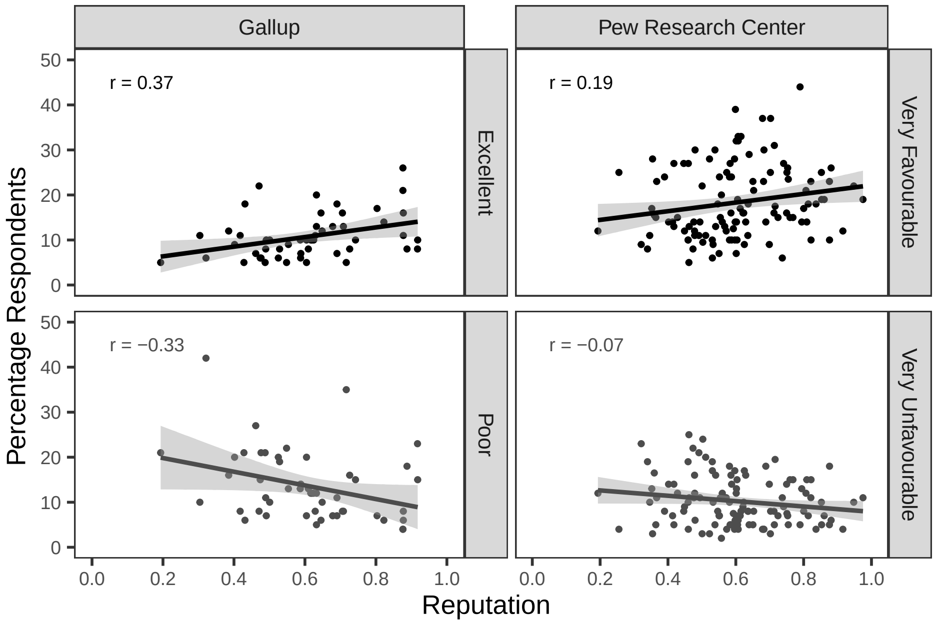

Figure 6: Reputation estimates of US federal agencies (x-axis) and percentage of respondents with an opinon about the agency as reported in the panel labels (y-axis). Correlation coefficients reported in each panel. Vertical lines are linear approximation with robust confidence intervals. Panels on the left-hand side use public opinion data from Gallup, whereas panels on the right-hand side use public opinion data from Pew Research Center.

The last validity test consists of comparing the reputation estimates with other estimates produced from an alternative corpus. Especially for more salient agencies – and therefore mentioned more often in the news – the estimates derived from speeches can be compare against estimates derived from newspaper articles. I estimated agency reputation from a corpus of more than 1.2 million articles published between 2014 and 2019 in 12 main UK national newspapers. Among the agencies most mentioned in the news – those with the lowest level of uncertainty – there is a positive and high correlation between the reputation estimates derived from the two corpora, thus strengthening the confidence in the criterion validity of the estimates.[9]

In the next section, I present an empirical application, showing how this measurement strategy can open new research agendas in the study of political-administrative interactions.

Bureaucratic Polarisation: An Application

Polarisation is a key characteristic of contemporary politics (Hacker and Pierson 2006). However, we do not know whether polarisation is an appropriate lens through which studying the bureaucracy. If reputation is ultimately a set of beliefs among audiences, does partisanship contribute to the formation of these beliefs?

One major advantage of measuring reputation with text-analysis methods is that, by meaningfully splitting the initial corpus, it is possible to break down the estimates by audiences. One way of doing it is to decompose the estimates along the partisan divide and measure the reputation agencies enjoy among different political parties (and arguably different audiences). Do agencies enjoy a different reputation among liberal and conservative parties? If so, how does this difference change over time and across agency type? To answer these questions, I replicated the estimation and trained the models on two different corpora, one for each of the two main parties in each country. The absolute value of the difference between the two partisan measures of reputation can thus represent a measure of "bureaucratic polarisation."

Figure 7 plots the average polarisation with respect to government/executive departments and non-departmental bodies over time. Surprisingly, polarisation is on average higher in the UK than in the US. Polarisation about government departments follows a slightly increasing trend in the US, whereas the trend is decreasing for independent

[9] In the SI (pp. 21-22) I report more information on the corpus of articles and the correlation between estimates.

agencies/non-departmental bodies, particularly in the 80s-90s. The trend for the UK is less clear. Polarisation about government departments increases during the Thatcher governments of the 80s and spikes again in the early 2000s with the Blair governments, while non-departmental bodies follow a decreasing trend.

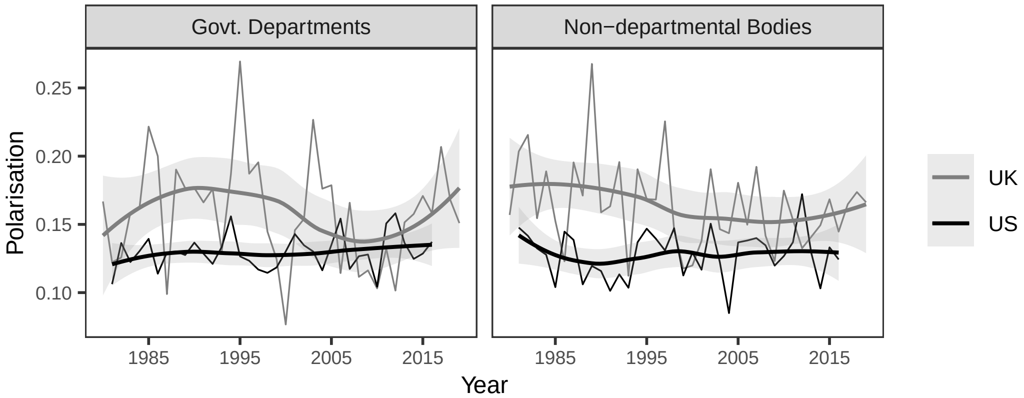

Figure 7: Average polarisation with respect to government departments and non-departmental bodies in the US and the UK. Each data point is the absolute difference of the average reputation among the two main political parties for all agencies in every year. For the US, government departments are executive departments, executive agencies, and agencies within the Executive Office of the President. Non-departmental bodies are independent agencies. Non-for-profit public organisations and government-owned corporations are excluded from the figure.

The varying level of politicisation of bureaucratic agencies may let us expect that their reputation depends on who leads the executive. Figure 8 shows the average reputation by party and by party in government (the label of the panels). In both countries, "partisan" reputation is higher when the party is in power compared to when it is at the opposition, with changes being particularly marked for liberal parties (+.04 and +.02 for the Labour and Democratic Party, respectively).

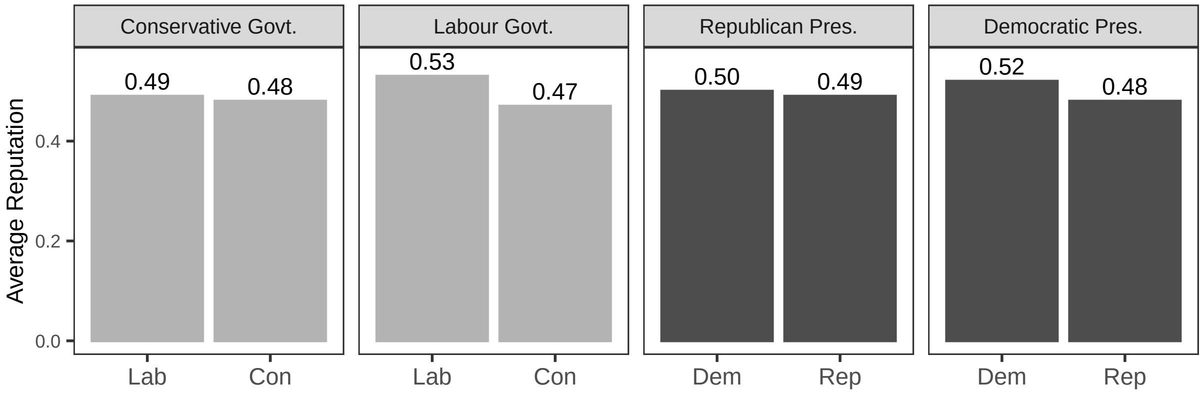

Figure 8: Average reputation by party (x-axis) and party in government (panel labels) for UK and US agencies. Y-axis reports the average reputation for every agency and across the entire period covered by the dataset.

The last layer to this snapshot of partisan measures of reputation is agency structure. Partisan differences might be more pronounced if the agency is under tight government control compared to more independent agencies. As a result, high levels of independence – and therefore low levels of politicisation – might be associated with lower bureaucratic polarisation.

To test this proposition more rigorously, I match the estimates of partisan reputation with the dataset on the structural independence of US federal agencies assembled by Selin (2015), which captures agency independence along two dimensions: independence as the ability of an agency to make policy decisions without political interference, and independence as statutory limitations to appointment/removal and qualification requirements placed on agency officials with key decision-making roles. The indicators are derived by modelling 50 structural features about the agencies with a Bayesian latent variable model. They range between 0 and 4, with higher values signifying higher independence. As shown in Table 3, independence is negatively associated with bureaucratic polarisation, although only with respect to the requirements placed on the officials who manage the agency. Models from (1) to (4) report OLS estimates, whereas Models from (5) to (8) report the results of WLS regressions, with weights equal to 1 over the average of uncertainty estimates across the two parties (see Section *Uncertainty*), so that most mentioned agencies are assigned larger weights. I also include agency type fixed effects, therefore accounting for the differences between departments, executive agencies, independent agencies, non-for-profit public organisations, and agencies within the Executive Office of the President. Far from causal interpretation, when pulling the regression towards matching the data with the lowest levels of uncertainty (WLS estimates), a one unit increase in independence as requirements imposed on agency officials is associated with a decrease in polarisation by .04, which is equal to a decrease by 29% with respect to the average polarisation across the sample (.13). The distance between bureaucratic reputation among Republicans and Democrats is smaller when appointment/removal limitations and qualification requirements exist on agency officials in key decision-making position.[10]

	OLS				WLS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ind: Decision-Makers	-0.01		-0.01	-0.02	-0.01		$\mathbf{-0.04}^{*}$	$\mathbf{-0.04}^{*}$
	(0.01)		(0.02)	(0.02)	(0.01)		(0.02)	(0.02)
Ind: Political Review		-0.00	0.01	0.01		0.01	0.02	0.01
		(0.01)	(0.01)	(0.01)		(0.01)	(0.01)	(0.01)
Indep. Agency (dummy)			-0.01				0.03	
			(0.03)				(0.02)	
Agency Type FE	No	No	No	Yes	No	No	No	Yes
R^2	0.01	0.00	0.01	0.03	0.00	0.01	0.06	0.08
Num. obs.	102	102	102	102	102	102	102	102

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: OLS and WLS estimates, standard errors in parenthesis. DV is distance between reputation estimates measured from speeches given by Republican and Democratic legislators. Independence data collected from agency statutes in 2013-2014, therefore the dependent variable is average reputation in 2013-2014. Agency type coded in the same way as listed on the institutional website of the US government, usa.gov/branches-of-government.

[10] In the SI (pp. 22-23) I show the estimates are robust to using heteroskedasticity-consistent standard errors, an alternative coding of the type of agencies, and to limiting the dataset to 2014, the year when the data collection in Selin (2015) ended.

Discussion and Limitations

Despite the significant advantages of a dynamic measure of reputation, it is worth emphasising some limitations of the proposed measurement strategy.

First, like every quantitative measure of agency attributes (e.g., discretion, autonomy, or politicisation), the proposed measure of reputation is a simplified picture of a conceptually rich and multifaceted attribute. Although word embeddings encode rich semantic features of terms, they are simply not able to match the deep observation of qualitative work. The construct validity of the measure could be enhanced by focusing on a smaller sample of agencies and making additional theoretically informed decisions about the estimation procedure. For instance, researchers could limit the textual corpus to a sub-set of speeches given by certain legislators or committee members, or about a pre-defined set of topics.

Second, and relatedly, modelling the total population of politicians' speeches might increase measurement error, for not everything said about bureaucracy should contribute to the reputation estimates. Let us consider speeches praising war veterans while at the same time mentioning the Department of Veterans Affairs. These speeches, arguably very positive, could inflate the reputation of the department even though they convey little information about its reputation. In the SI (pp. 18-19) I show that the reputation of US military agencies decreases when speeches mentioning both the name of military agencies and military values (e.g., integrity, honour, courage, etc.) are excluded from the corpus. However, while removing speeches mentioning military agencies and military values might "de-bias" the estimates, we could in fact be cancelling out an important dimension of their reputation, namely the fact that they operate in a salient and respected political domain. Decisions about the textual corpus should therefore be driven by strong theoretical reasons. Other sources of heterogeneity at the agency level that cannot be handled by sampling speeches can be addressed empirically too. For instance, looking at changes in reputation over time can account for the different (and constant) probability of agencies being mentioned alongside more positive concepts.

Third, politicians' speeches – while giving voice to agencies' audiences – might also include an undesired partisan component. This can be problematic for the reputation estimates of more politicised agencies, which may be subject to more partisan debates. As shown in the application, there seem to be partisan differences in the way politicians talk about bureaucracy. Yet the partisan divide is just one out of the many ways we could split the speeches. Other differences could be detected based on the legislative roles, education, or previous political experience of politicians. While the partisan divide is the most intuitive way to group legislative speeches, future research could study other cleavages too, and use them to explain how audiences form their beliefs about bureaucracy as well as audiences' behaviour *vis-à-vis* administrative agencies.

Similarly, while parliamentary speeches are a rich source of information about agencies' reputation, they might not fully or precisely capture key agencies' audiences. Neglecting some audiences or mischaracterising their beliefs would bias the measure of reputation, which would result in reputation estimates diverging the agency's "true" reputation. While this might be a source of concern for researchers willing to use this measure, it can also represent an opportunity to investigate how reputation is portrayed across different venues. As I show in the SI (pp. 21-22), for highly salient agencies, newspaper articles might be a good alternative and complementary source of information about agencies' reputation.

Finally, despite the new opportunities for quantitative analysis offered by this dynamic measure of reputation, comparative research should be guided by robust theory, especially when defining the sample of agencies to study. As argued by D. Carpenter (2020), comparing administrative bodies across policy domains is often likely to lead to flawed conclusions. This is particularly relevant for the study of reputation, which underscores the importance of agency's reputation for uniqueness, and hence focuses on agencies' reputation *within* their own field.[11] Depending on the research question at hand, looking at within-agency variation might provide a solution to the "comparative trap" if the features of the policy domain remain constant over time, but researchers should be careful when reaching conclusions based on comparative analyses that pool many agencies with different tasks, missions, and organisational features.

Conclusions

Structural accounts of bureaucratic politics have long claimed that formal decisions about the agency are so incisive that they allow politicians to define the balance between agency discretion and political control. Yet on the other hand, it is thanks to reputation that agencies accrue their ability to "sway the wishes of elected officials on particular matters of policy and to secure deference from these elected officials" (D. P. Carpenter and Krause 2012, 30). Bureaucratic reputation has been used to study important bureaucratic outcomes, from strategic communication, to autonomy, and accountability, yet empirical work has not been able to match the rich theoretical and conceptual ground of reputational theories of bureaucracy, partly because we lack a dynamic measure of reputation.

I first claimed that bureaucratic reputation can be aptly captured from what politicians say, for they have electoral incentives to align with key agencies' constituencies. By employing word-embedding techniques, I estimated the reputation of 465 bureaucratic bodies from more than 4.5 million speeches over a period of almost 40 years in two

[11] In the SI (p. 17) I give two examples of within-domain comparisons, looking at the reputation of US agencies overseeing financial institutions and UK independent regulators of network industries.

major western democracies, the US and the UK. I performed multiple validation tests and presented an application of this method to the study of bureaucratic polarisation, opening new research agendas that combine key concepts in bureaucratic politics such as agency structure and politicisation with theories of partisan identity and polarisation. In particular, future research could re-answer classical questions in bureaucratic politics by measuring how the degree of polarisation in principals' beliefs about the agent affects the interactions between politicians and bureaucracy.

This paper makes two main contributions. One methodological, showing how recent advances in natural language processing can be employed to study bureaucratic attributes and political-administrative interactions; one substantive, offering an original and validated measure of reputation able to capture variation across agencies and over time. Overall, this measure will be able to advance our understanding of key questions in administrative politics, from the causes and effects of bureaucratic reputation, to more nuanced questions about delegation of authority, political control, and bureaucratic accountability, while simultaneously opening new uncharted research agendas that bridge bureaucratic politics with other subfields of political science.

References

Anastasopoulos, Jason, and Anthony M. Bertelli. 2019. "Understanding Delegation through Machine Learning: A Method and Application to the European Union." *American Political Science Review*, 1–11. https://doi.org/10.2139/ssrn.3207821.

Anastasopoulos, L Jason, and Andrew B Whitford. 2019. "Machine Learning for Public Administration Research, With Application to Organizational Reputation." *Journal of Public Administration Research and Theory* 29 (3): 491–510. https://doi.org/10.1093/jopart/muy060.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data." *American Political Science Review* 113 (4): 883–901. https://doi.org/10.1017/S0003055419000352.

Bertelli, Anthony M, Dyana P Mason, Jennifer M Connolly, and David A Gastwirth. 2013. "Measuring Agency Attributes with Attitudes Across Time : A Method and Examples Using Large-Scale Federal Surveys." *Journal of Public Administration Research and Theory* 25: 513–44. https://doi.org/10.1093/jopart/mut040.

Blumenau, Jack. 2019. "The Effects of Female Leadership on Women's Voice in Political Debate." *British Journal of Political Science*, 1–22. https://doi.org/10.1017/s0007123419000334.

———. 2021. "House of Commons Parliamentary Debates, 1979-2019." Colchester, Essex: UK Data Service.

Busuioc, E. Madalina. 2016. "Friend or foe? inter-agency cooperation, organizational reputation, and turf." *Public Administration* 94 (1): 40–56. https://doi.org/10.1111/

padm.12160.

Busuioc, E. Madalina, and Martin Lodge. 2016. "The Reputational Basis of Public Accountability." *Governance* 29 (2): 247–63. https://doi.org/10.1111/gove.12161.

Busuioc, Madalina, and Dovilė Rimkutė. 2019. "Meeting expectations in the EU regulatory state? Regulatory communications amid conflicting institutional demands." *Journal of European Public Policy*. https://doi.org/10.1080/13501763.2019.1603248.

Carpenter, Daniel. 2020. "On Categories and the Countability of Things Bureaucratic: Turning From Wilson (Back) to Interpretation." *Perspectives on Public Management and Governance* 3 (2): 83–93. https://doi.org/10.1093/ppmgov/gvz025.

Carpenter, Daniel P. 2001. *The Forging of Bureaucratic Autonomy.* Princeton: Princeton University Press.

———. 2002. "Groups, the Media, Agency Waiting Costs, and FDA Drug Approval." *American Journal of Political Science* 46 (3): 490–505. https://doi.org/10.2307/3088394.

———. 2010. *Reputation and Power. Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton: Princeton University Press.

Carpenter, Daniel P., and George A. Krause. 2012. "Reputation and public administration." *Public Administration Review* 72 (1): 26–33.

———. 2015. "Transactional authority and bureaucratic politics." *Journal of Public Administration Research and Theory* 25 (1): 5–25. https://doi.org/10.1093/jopart/muu012.

CNN. 2014. "Bad VA care may have killed more than 1,000 veterans, senator's report says." edition.cnn.com/2014/06/24/us/senator-va-report/index.html.

Collier, David, and Robert Adcock. 2001. "Measurement Validity: A Shared Standard for

Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–46.

Epstein, David, and Sharyn O'Halloran. 1999. *Delegating Powers: A Transaction Cost Politics Approach to Policy Making under Separate Powers.* Cambridge, UK: Cambridge University Press.

Firth, John Rupert. 1957. *Studies in linguistic analysis.* Wiley-Blackwell.

FSA. 2009. "The Turner Review. A regulatory response to the global banking crisis." March. papers2://publication/uuid/3CA20D05-3397-4609-92CC-40C8E0AB881E.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences of the United States of America* 115 (16): E3635–44. https://doi.org/10.1073/pnas.1720347115.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2018. "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts." Palo Alto, CA: Stanford Libraries [distributor]. https://data.stanford.edu/congress%7B/_%7Dtext.

Gilad, Sharon, Moshe Maor, and Pazit Ben Nun Bloom. 2013. "Organizational reputation, the content of public allegations, and regulatory communication." *Journal of Public Administration Research and Theory* 25 (2): 451–78. https://doi.org/10.1093/jopart/mut041.

Gilad, Sharon, and Tamar Yogev. 2012. "How Reputation Regulates Regulators: Illustrations from the Regulation of Retail Finance." In *The Oxford Handbook of Corporate Reputation*, edited by Timothy G. Pollock and Michael L. Barnett, 1–24. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199596706.013.0016.

Hacker, Jacob S., and Paul Pierson. 2006. *Off Center: The Republican Revolution and* *the Erosion of American Democracy*. New Haven, CT: Yale University Press.

Hood, Christopher. 1991. "A Public Management For All Seasons?" *Public Administration* 69: 3–19. https://doi.org/10.2307/25305228.

House of Commons Defence Committee. 2011. "The Strategic Defence and Security Review and the National Security Strategy." July.

Huber, John D., and Charles R. Shipan. 2002. *Deliberate Discretion? The institutional foundations of bureaucratic autonomy*. New York: Cambridge University Press.

Krause, George A., and J. Kevin Corder. 2007. "Explaining bureaucratic optimism: Theory and evidence from U.S. executive agency macroeconomic forecasts." *American Political Science Review* 101 (1): 129–42. https://doi.org/10.1017/S0003055407070074.

Krause, George A., and James W. Douglas. 2005. "Institutional design versus reputational effects on bureaucratic performance: Evidence from U.S. government macroeconomic and fiscal projections." *Journal of Public Administration Research and Theory* 15 (2): 281–306. https://doi.org/10.1093/jopart/mui038.

Lee, Soo Young, and Andrew B. Whitford. 2013. "Assessing the effects of organizational resources on public agency performance: Evidence from the US federal government." *Journal of Public Administration Research and Theory* 23 (3): 687–712. https://doi.org/10.1093/jopart/mus050.

Lewis, David E. 2008. *The Politics of Presidential Appointments: Political Control and Bureaucratic Performance*. Princeton University Press.

Lowande, Kenneth. 2018. "Who Polices the Administrative State?" *American Political Science Review* 112 (4): 874–90. https://doi.org/10.1017/S0003055418000497.

Maor, Moshe. 2007. "A scientific standard and an agency's legal independence: Which of these reputation protection mechanisms is less susceptible to political moves?" *Public*

Administration 85 (4): 961–78. https://doi.org/10.1111/j.1467-9299.2007.00676.x.

Maor, Moshe, Sharon Gilad, and Pazit Ben Nun Bloom. 2013. "Organizational reputation, regulatory talk, and strategic silence." *Journal of Public Administration Research and Theory* 23 (3): 581–608. https://doi.org/10.1093/jopart/mus047.

Maor, Moshe, and Raanan Sulitzeanu-Kenan. 2013. "The Effect of Salient Reputational Threats on the Pace of FDA Enforcement." *Governance* 26 (1): 31–61. https://doi.org/10.1111/j.1468-0491.2012.01601.x.

Maor, Moshe, and Arild Wæraas. 2015. "Understanding organizational reputation in a public sector context." In *Organizational Reputation in the Public Sector*, edited by Moshe Maor and Arild Wæraas, 1st ed. New York: Routledge. https://doi.org/https://doi-org.libproxy.ucl.ac.uk/10.4324/9781315850825.

McCubbins, M, Roger G Noll, and Barry R Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics, & Organization* 3 (2): 243–77.

Ministry of Defence. 2010. "Securing Britain in an Age of Uncertainty: The Strategic Defence and Security Review Securing Britain in an."

———. 2011. "Defence Reform. An independent report into structure and management of the Ministry of Defence." June. https://doi.org/10.1080/1356788000642.

Moe, Terry M. 1990. "Political Institutions: The Neglected Side of the Story." *Journal of Law, Economics, & Organization* 6: 213–53.

Nelson, David, and Susan Webb Yackee. 2012. "Lobbying coalitions and government policy change: An analysis of federal agency rulemaking." *Journal of Politics* 74 (2): 339–53. https://doi.org/10.1017/S0022381611001599.

Office of Inspector General (DHS). 2015. "Evaluation of DHS ' Information Security

Program for Fiscal Year 2015." oig.dhs.gov/assets/Mgmt/2016/OIG-16-08-Nov15.pdf.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43.

Pew Research Center. 2019. "Public Expresses Favourable Views of a Number of Federal Agencies." https://doi.org/.1037//0033-2909.I26.1.78.

Potter, Rachel Augustine. 2017. "Slow-rolling, fast-Tracking, and the pace of bureaucratic decisions in rulemaking." *Journal of Politics* 79 (3): 841–55. https://doi.org/10.1086/690614.

Preotiuc-Pietro, Daniel, Daniel J. Hopkins, Ye Liu, and Lyle Ungar. 2017. "Beyond binary labels: Political ideology prediction of twitter users." *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1: 729–40. https://doi.org/10.18653/v1/P17-1068.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R Radev. 2010. "How to Analyze Political Attention.pdf." *American Journal ofPolitical Science* 54 (1): 209–28.

Rheault, Ludovic, and Christopher Cochrane. 2019. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis*, 1–22. https://doi.org/10.1017/pan.2019.26.

Rice, Douglas R., and Christopher Zorn. 2019. "Corpus-based dictionaries for sentiment analysis of specialized vocabularies." *Political Science Research and Methods*, 1–17. https://doi.org/10.1017/psrm.2019.10.

Roberts, Patrick S. 2006. "FEMA and the prospects for reputation-based autonomy." *Studies in American Political Development* 20 (1): 57–87. https://doi.org/10.1017/S0

898588X06000010.

Selin, Jennifer L. 2015. "What Makes an Agency Independent?" *American Journal of Political Science* 59 (4): 971–87. https://doi.org/10.1111/ajps.12161.

Spirling, Arthur. 2016. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832-1915." *Journal of Politics* 78 (1): 120–36. https://doi.org/10.1086/683612.

Spirling, Arthur, and Pedro L Rodriguez. 2019. "Word Embeddings What works, what doesn't, and how to tell the difference for applied research."

The Daily Telegraph. 2008. "FSA staff will get bonuses despite overseeing financial crisis." 3. telegraph.co.uk/news/3374017/FSA-staff-will-get-bonuses-despite-overseeing-financial-crisis.html.

The Guardian. 2009. "Jacqui Smith goes - hit by porn claims, but fatally damaged by 42 days." theguardian.com/politics/2009/jun/02/jacqui-smith-resignation.

The Times. 2018. "Home Office knew Windrush scandal was brewing for years." thetimes.co.uk/article/home-office-knew-windrush-scandal-was-brewing-for-years-ws08ghmxx.

The Washington Post. 2008a. "DHS Withdraws Bid to Curb Union Rights." washingtonpost.com/wp-dyn/content/article/2008/02/19/AR2008021902459{_}pf.html.

———. 2008b. "Congress Says DHS Oversaw $15 Billion in Failed Contracts." washingtonpost.com/wp-dyn/content/article/2008/09/16/AR2008091603200{_}pf.html.

Thomson, Robert, and René Torenvlied. 2011. "Information, Commitment and Consensus: A Comparison of Three Perspectives on Delegation in the European Union." *British Journal of Political Science* 41 (1): 139–59. https://doi.org/10.1017/S0007123410000268.

Timeline. 2017. “Snapshots from FEMA’s disastrous failure in the wake of Hurricane Andrew.” timeline.com/fema-hurricane-andrew-snapshots-7a764f017614.

Vannoni, Matia, Elliott Ash, and Massimo Morelli. 2020. “Measuring Discretion and Delegation in Legislative Texts: Methods and Application to US States.” *Political Analysis.*

Wood, Matthew, Sjors Overman, and Madalina Busuioc. 2020. “A Multidimensional Reputation Barometer for Public Agencies: A Validated Instrument.” *Public Administration Review,* 1–11. https://doi.org/10.1111/puar.13158.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511818691.