Measuring and Explaining Political Sophistication Through Textual Complexity

Running Title: Measuring Textual Complexity[*]

Kenneth Benoit[†] Kevin Munger[‡] Arthur Spirling[§]

October 2, 2018

Keywords: text analysis, political communication, readability, sophistication

Word Count: 9,614 words

[*] This research was partly supported by the European Research Council grant ERC- 2011-StG283794-QUANTESS.

[†] Professor of Quantitative Social Research Methods, London School of Economics (`kbenoit@lse.ac.uk`)

[‡] PhD Candidate, Department of Politics, New York University (`km2713@nyu.edu`)

[§] Associate Professor of Politics and Data Science, New York University (`arthur.spirling@nyu.edu`)

Abstract

Political scientists lack domain-specific measures for the purpose of measuring the sophistication of political communication. We systematically review the shortcomings of existing approaches, before developing a new and better method along with software tools to apply it. We use crowdsourcing to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of sophistication. This includes previously excluded features such as parts of speech and a measure of word rarity derived from dynamic term frequencies in the Google books dataset. Our technique not only shows which features are appropriate to the political domain and how, but also provides a measure easily applied and re-scaled to political texts in a way that facilitates probabilistic comparisons. We reanalyze the State of the Union corpus to demonstrate how conclusions differ when using our improved approach, including the ability to compare complexity as a function of covariates.

Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: `https://doi.org/10.7910/DVN/9SF3TI`.

1 Introduction

A key concern in the study of politics is how the nature of political communication has changed. At the same time that the challenges of governing have grown in complexity, the sophistication of political speech, by many measures, appears to have declined. Typically as part of a broader discussion concerning "dumbing down" (Gatto, 2002), scholars have applied measures of textual complexity from educational fields to find that the sophistication of political language has steadily decreased over the past 200 years (e.g. Lim, 2008). Such concerns are echoed in popular presentations, and it is not uncommon to see media analysis assessing political speeches in terms of the (purported lower) school grade level required to understand them.[1]

By contrast, and with more optimistic conclusions, other social science studies have used measures of textual complexity to link linguistic sophistication to outcomes, with a focus on the concrete benefits to clarity. Jansen (2011), for instance, studies the reading level of communications from four central banks, equating lower reading levels of bank communication with greater clarity, which they link to positive effects on the volatility of financial market returns. Likewise, Owens and Wedeking (2011) and Spriggs (1996) examine the complexity of Supreme Court decisions, pointing to the importance of clarity in court opinions. In the context of the British parliament, Spirling (2016) applies readability measures to document the democratizing effects of franchise reform on elite speeches. Studying post-war Austrian and German elections, Bischof and Senninger (2018) find that simpler manifestos make for better informed voters. Finally, as a meta-analysis to defend against charges of elitism and jargon (e.g. Diamond, 2002; Kristof, 2014), Cann, Goelzhauser and Johnson (2014) show that while the reading ease of articles in the top political science journals has declined since 1910, the typical political science article requires less reading ability than the average article in *Time Magazine* or *Reader's Digest*.

These applications share one trait: they equate important substantive characteristics of politi-

[1] For instance: "Trump Speaks at Fourth-grade Level, Lowest of Last 15 U.S. Presidents, New Analysis Finds," *Newsweek*, January 8, 2018, `http://www.newsweek.com/trump-fire-and-fury-smart-genius-obama-774169`. See also: *The Guardian*, 12 February 2013, `http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level`

cal, economic, or legal communication—such as clarity or sophistication—with indexes such as the Flesch Reading Ease (FRE) score (Flesch, 1948). These measures, however, were developed decades earlier in entirely different contexts, namely educational research and applied psychology, and their applicability to contemporary political speech remains untested. Consequently, we are uncertain as to the true direction of change for specifically *political* communication. More importantly perhaps, because our current measurement strategies are weak, we find it hard to disentangle changes to texts which are normatively positive ("clearer") versus negative ("dumbing down"). For example, the fact that we might communicate the same complex content, but in shorter words and sentences that require less processing effort by the reader, is almost certainly a good thing. Yet, as we will see, traditional measures imply such changes are in line with appealing to a less educated audience and thus deemed a source of concern.

To address such problems, here we systematically review the properties and statistical performance of current measures of textual difficulty, and develop a new measure designed specifically for political language. Our approach uses experimental data based on human pairwise comparisons of short extracts of political speech (e.g. Lowe and Benoit, 2013; Montgomery and Carlson, 2017), which we then use to scale linguistic sophistication using a scaling approach developed by Bradley and Terry (1952) to measure latent "ability" from pairwise contests, treating the reading ease of a text as equivalent to ability. This approach permits more direct statements about uncertainty and inference, including the *probability* that a given text is easier or harder, either to one another or to a known benchmark, such as a fifth-grade reading level. Rather than relying on static estimates fit to data from a non-political context, our approach allows flexible determination of the components of textual sophistication, as well as their appropriate weights, in a manner that can be adapted to any domain but that is fit here to texts from the US State of the Union (SOTU) corpus to provide a specifically political measure of textual sophistication. Generalizing beyond this corpus, our contribution is to set out clear principles for measuring linguistic sophistication in the political domain, demonstrate the methodological superiority of our approach, and to outline a new method for fitting appropriate measures to any context.

2 Measuring the Sophistication of Political Communication

We first define terms and review previous efforts.

2.1 Textual Sophistication, Complexity, and Difficulty

As applied to text, we use the terms “sophistication”, “difficulty” and “complexity” somewhat interchangeably, reflecting ambiguity in existing use of these terms.

For Luskin (1990), to give a political communication example, *sophistication* is a property of individuals rather than messages, and pertains to how elaborate is the individuals’ political belief system. Thus, “[a] person is politically sophisticated to the extent to which his or her political cognitions are numerous, cut a wide substantive swath, and are highly organized, or ‘constrained”’ (Luskin, 1990, 332). In that world, measurement focuses on the interest that a citizen has in politics, her educational level, and exposure to current events and related variables. Of course, this conception does not lend itself naturally to a measure for texts themselves. For example, it is unclear whether a document written in a simple way about an obscure (but wholly political) issue ought to be considered more or less sophisticated than one written about a well-known subject that requires some prior education to appreciate fully.

In linguistics, *complexity* is a characteristic of a text, but there are multiple measures and thus multiple implied definitions in practice. For example, social scientists might well agree with Gibson (1998, 2) that complexity is the “quantity of computational resources . . . [that documents] require to process.” But they might disagree with his focus on ambiguity as to whether a transitive verb refers to the object or subject, the presence of particular types of nested clauses and the distance between certain elements in sentences. That is, a “sophisticated” political sentence is not merely confusing or hard to follow.

Perhaps the simplest way to conceptualize the measurement problem comes from education research, where the concern is to match learning materials to students based on their age and cognitive ability (see Klare, 1963, for an overview). There, the emphasis is on the “readability” of a document, and the intuitive notion that one text may be relatively more "difficult" than another in terms of some downstream comprehension task (e.g. a school test about the passage or book in question). In this vein, textual difficulty embodies some mix of the concepts above. If the message in a text is subtle, and can only really be understood or appreciated by a well-educated person, it is both difficult *and* sophisticated.[2] Meanwhile, if a document is written with an unusual or archaic (but nonetheless correct) grammatical structure, it is both difficult *and* complex.[3] While these concepts are not exactly equivalent, their ready application to texts based on empirically established markers has encouraged their widespread adoption in educational research to measure the reading difficulty of texts, a usage whose application has spread to other fields. For this reason, and because these formulations are so straightforward, we focus our efforts here on improving these measures as they apply to political texts.

2.2 Traditional Measures of Textual Difficulty

Measuring the difficulty of educational texts is not new (e.g. Sherman, 1893), and there are now a large number of indexes for this task—indeed, michalke (2017) references and implements no fewer than 27 of them—but the various Flesch-based metrics (Flesch, 1948, 1949; Kincaid et al., 1975) have dominated.

In terms of technical details, for a given document, the traditional measures of reading difficulty take into account some combination of: (average) sentence length (e.g. Flesch, 1948, 1949; Gunning, 1952; Fry, 1968; Kincaid et al., 1975); the (average) number of syllables per word (e.g. Flesch, 1948, 1949; Gunning, 1952; Wheeler and Smith, 1954; Fry, 1968; Kincaid et al., 1975); the parts of speech represented in the document (e.g. Coleman and Liau, 1975); and the (average) familiarity of the terms used (e.g. Dale and Chall, 1948; Spache, 1953).

Flesch's (1948) pioneering work focused on the reading comprehension of school children:

[2] But perhaps not complex for them: for example, a statistics text might use the terms *moment* and *distribution* in a way that is not ambiguous to a political methodologist.

[3] But perhaps not sophisticated: for example, reading a non-commissioned officer's diary from the American Civil War might be hard work for a modern reader, but not because it is discussing abstruse themes.

in particular, the average grade of students who could correctly answer at least 75% of some multiple choice questions regarding a few select texts. This dependent variable was subsequently transformed to a zero to 100 scale, and regressed on a constant and two predictors (average sentence length and average number of syllables per word). This yielded the following formula for scoring documents:

$$206.835 - 1.015\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right) - 84.6\left(\frac{\text{total number of syllables}}{\text{total number of words}}\right).$$

Known as the *Flesch Reading Ease* score, this measure had the intended range "for almost all samples taken from ordinary prose" (Flesch, 1948, 225).[4] Subsequently, Kincaid et al. (1975) introduced a mechanical conversion of the formula that yields values roughly equivalent to the US grade school level required to understand a text.

Other than indirectly through syllable counts, the Flesch formula does not take into account the actual familiarity of the words used in a text. An example of an approach that does is Dale-Chall (Dale and Chall, 1948), whose key difference from the Flesch index was its replacement of the word length input by a text's percentage of "difficult" words, specified as any word not included in a list 763 words deemed to be those known by 80% of fourth grade children (in 1948).[5]

3 Improving measures of textual sophistication

While political scientists have not ignored the measurement of readability (e.g. Cann, Goelzhauser and Johnson, 2014), there has not been especially great interest in adapting them to specifically political contexts. This gives rise to two broad sets of issues that give us pause: first, *theory-based* concerns related to what using such measures implies about the elements that determine textual sophistication and their appropriate weights; and second, a general lack of desirability from a

[4] In practice, the statistic is bounded at an upper "ease" limit of 121.22 for texts consisting of one-syllable, one-word sentences, and bounded from below only by an offset of the average word and sentence length.

[5] This was later expanded to around 3,000 words. The formula has also been adjusted over time (Chall and Dale, 1995), but originally was: $0.1579\,(\text{percentage of difficult words}) + 0.0496\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right)$.

statistical perspective.

3.1 Empirically determining the indicators of textual sophistication

Traditional measures of readability use different combinations of indicators and weights. Since the 1970s, however, such measures have been criticized as atheoretical, at least in terms of the way that educational researchers thought about cognition (Kintsch and Vipond, 1979). Consequently, scholars have treated them with increasing caution because their performance was found wanting in a series of studies (e.g. Bruce, Rubin and Starr, 1981; Smith, 1986). Since none of those contexts was political, furthermore, the arbitrary choice of indicators from studies in non-political domains makes the case for fitting a specifically political measure of textual sophistication all the more compelling. In particular, the schoolchildren studied in most previous approaches may not be representative of the adult citizens we care about for political science cases. And while what makes a *political* text difficult may be somewhat similar to the factors that make educational passages harder, this remains an empirical question to be examined. For a statistical model of textual easiness, such as the one we develop below, this means fitting the model to a large set of potential determinants of sophistication, within the context of domain-relevant texts. We now lay out our priors about what will matter, and why.

First, we expect *greater use of longer words* to indicate a higher degree of sophistication. As in education, longer words are assumed to make things harder for political audiences, whether this length occurs in the form of characters or syllables. Use of the noun *plebiscite*, for instance, signals greater textual sophistication than use of its synonym *vote*. Because political text is usually designed explicitly to deliver an ideological message, such deliberate choices may matter even more than similar indicators in school texts, where the goal is to educate and entertain.

Next, we expect that *greater use of relatively uncommon words* will indicate higher sophistication than use of their more commonplace synonyms. Not only do relatively rare words require a larger vocabulary, and hence a more widely read and more literate audience, but rarer words typically mark more precise, domain-specific language that is the hallmark of expertise. In political text, this can translate into more sophisticated content, as well as style.

Traditional measures of readability have captured word rarity in a static fashion, in the form of lists of "easy words" (e.g. Dale and Chall, 1948) or some difficulty measure attached to each word (e.g. Bonsall et al., 2017).

Word rarity is not static, however, especially with respect to changing lexicons over time. The term *husbandry* (the cultivation and breeding of crops and animals) was used much more often in the 1790s than in current times, and therefore its inclusion in a list of easy or difficult words today may be misleading for the prior period. Thus, we need to model contemporary understandings differently from more historical ones.

Longer sentences also reflect greater sophistication, whether measured in words or characters, since these not only reflect more complex ideas but also require more attention to absorb, in the linguistic sense we mentioned above. For this reason, nearly every previous measure of reading difficulty takes sentence length into account in some form.

Finally, more sentences with *more complex syntactic and grammatical structures* indicate greater sophistication. Beyond length, structural complexity in the form of multiple or subordinate clauses indicates that more complex ideas are being communicated. This may also take the form of greater reliance on particular parts of speech, such as nouns or adjectives. We know that politicians use stories or anecdotes as a rhetorical device to exemplify a given policy or reform (Charteris-Black, 2011). In this light, we can imagine that, per Flesch (1948), more "compelling" political texts—that invoke *human interest* via noun usage (over other parts of speech)–are deemed easier to understand. Such content should be modelled, but presumably was not in previous measures due primarily to the lack of reliable, automatic natural-language processing (NLP) tools to parse dependency structures or tag grammar. In our application below, we use modern techniques in this area to capture the role of both grammar and syntax as it affects textual sophistication.

To capture the varieties of these potential determinants of political sophistication and to determine their appropriate weights for our context, we include 22 possible indicators (described below). As with all previous investigations into the appropriate indicators of reading difficulty, we do not purport to outline a full human linguistic model of the "data-generating process" of textual sophistication. However, our approach is able to consider a comprehensive set of domain-specific candidate inputs, including the fixed set or fixed weights of those used in traditional measures. We leave the weights of each input's contribution as an empirical question to be tested in context, and not one whose answer can be determined from theory or from findings derived in different settings.

3.2 Improving the statistical properties of sophistication measures

Even if we have managed to fit a "correct" set of variables to measure textual sophistication, statistical issues remain. Traditional measures are simply weighted sums—they are not fit to anything other than the original data, and so do not maximize a well-defined objective function. The immediate consequence is that we cannot know whether a given measure is performing well or not, statistically, on new texts. And thus, we *cannot naturally compare* different measures on the same data. Perhaps unsurprisingly given the lack of an objective function, there are also *no uncertainty* estimates associated with document scores. Yet surely (in the sense of Lowe and Benoit, 2013) we think that, holding the indicators constant, a text with greater values of indicators related positively to sophistication provides more evidence of a given level of (latent) difficulty than a text with lower values on those indicators. Finally, using a static index approach means that fine-grained differences in scores have essentially no useful interpretation. There are two elements to this issue: first, continuous estimates from measures like FRE only really apply to the original school children in Flesch's study. In light of this, it is unclear what it means to say one State of the Union speech is a "70" and another is a "75" in the year 2018. Second, there is no way to convert numbers like 70 and 75 into a framework that allows *probabilistic* inference—we mean this both in terms of the interpretation of the point estimates and in terms of the confidence intervals around them.

In what follows, we address this problem by providing an approach generated from pairwise comparisons and ideally suited to direct, probabilistic comparisons of difficulty, either between two texts or between a text and a known baseline. We demonstrate that this key feature, combined with uncertainty estimates, provides a far more useful comparative measure for political and social science than previous approaches.

4 Methodology for Fitting a Domain-Specific Measure of Textual Sophistication

We have two broad sets of problems to solve: first, determining the appropriate inputs, and their weights, for a model of textual sophistication that fits the political context better than the simple mechanical formulas derived from education research; and second, formulating this in an explicitly statistical framework that enables the direct, probabilistic statements needed for social scientific measurement and comparison.

Our workflow involves the following steps:

1. Get human judgments of relative textual easiness for specifically political texts.
   (a) Sample pairs of short, appropriate text segments ("snippets") that form a minimally connected set.
   (b) Get large numbers of human judgments as to which text segment is easier for each pair, using crowdsourcing.[6]
2. Fit an unstructured Bradley-Terry (Bradley and Terry, 1952) model for pairwise comparisons to the judgment data from Step 1, in order to estimate a measure of latent "easiness" as equivalent to the "ability" parameter in the Bradley-Terry framework.
3. Using the set of potential determinants of relative textual easiness estimate the best predictors of the textual easiness from Step 2 using the random forests algorithm.
4. Using only the most highly predictive variables from Step 3, fit a *structured* Bradley-Terry using the data from Step 1.
5. Use the fitted model from Step 4 to "predict" the easiness parameter for a given new text, including:
   (a) Using the comparative formulation to estimate the relative probability that one new text is easier than another text, or a baseline text; and
   (b) Using non-parametric bootstrapping of the new texts to represent uncertainty in the predicted point estimates.

[6] Because our pre-tests indicated that it was more straightforward to ask raters which text was *easier*, our subsequent discussion is about relative easiness rather than difficulty (similar to original Flesch scale in which higher values indicated easier texts).

Step 5 is similar to having re-engineered a classical difficulty measure, but with improved properties as a statistical estimator. In software to accompany this paper, we provide this fitted model along with functions to apply it to any new text. But by detailing the earlier steps, we not only provide full transparency as to how the new measure was produced, but also a reproducible workflow to enable this approach to be fit to new contexts. In the remainder of this section, we detail each of these steps.

4.1 Obtaining human judgments of relative textual easiness

Our measurement assumption is that human interpretation provides the "gold standard" for judging the relative sophistication of political text (or any text). Because there is no absolute metric of textual difficulty, we view this as a fundamentally comparative problem: What factors make one text more sophisticated than another? Our first step was to produce data consisting of roughly comparable, short segments of text, drawn from the political domain of interest, and obtain large numbers of human judgments as to which was easier than the other.

For data, we extracted a series of short texts of one or two sentences each—"snippets"—to be given to human coders to compare, pairwise. The coders tell us which of the two snippets is easier to understand, and they do this multiple times for various combinations of different snippets. In principle, we could have had the coders rate each snippet on some predefined scale, but experience demonstrates that humans find it considerably easier to do pairwise comparisons with respect to a trait (Thurstone, 1927; Montgomery and Carlson, 2017). Snippets are only segments of the original documents, of course, but asking raters to compare entire documents is infeasible. In addition, previous work based on coding document components (e.g. Benoit et al., 2016) indicates that, especially where the segments are of comparable length, this approach works well for recovering document characteristics.

To obtain the pairwise judgments, we recruited large numbers of non-experts to provide judgments in a fast, reproducible manner (e.g. Benoit et al., 2016) using a crowdsourcing platform. *Crowdsourcing* is a means of getting a large-scale task completed by dividing it into many small pieces and outsourcing the pieces in random order to a distributed, anonymous worker pool known as the "crowd." By reassembling the returned micro-tasks, the overall job is completed quickly and inexpensively by a pool of workers whose effort and attention adapt flexibly to the job requirements and their own willingness and availability. Our tasks used the *Figure Eight* crowd-sourcing platform.[7] The task was labeled as "Identify Which of Two Text Segments Contains Easier Language." Upon accepting the task, we provided the workers with a number of example comparisons, with one option correctly labeled as more complex, as well as a qualifying test of ten questions from which six had to be answered correctly in order to qualify for production tasks. The specific instructions provided to each worker were:

> Your task is to read two short passages of text, and to judge which you think would be easier for a native English speaker to read and understand. An easier text is one that takes a reader less time to comprehend fully, requires less re-reading, and can be more easily understood by someone with a lower level of education and language ability.

The snippets served for comparison were two-sentence segments drawn from the 70 State of the Union Addresses (SOTUs) delivered after 1950. We used these texts because the purpose of the SOTU addresses has remained relatively unchanged in the postwar period, and because of the attention these speeches have received in previous examinations of readability.

Some preprocessing of the addresses prior to creating snippets was required; we removed some organizational non-sentence pieces of text (mostly referring to the medium by which the address was delivered). Once cut down for comparison, we disqualified some snippets from consideration. We dropped those which were outside the 0–121 range of the FRE: this was a simple way to remove unusual texts that were much harder than an adult reader would typically encounter (note that 121 is the maximum easiness possible for the Flesch scale). We also removed snippets that contained more than two numeric years, had large numbers, or began with the title of a document section. These restrictions were put in place to avoid comparisons being made on dimensions which are not strictly connected to the "regular" textual content of a message.

We constrained the snippets drawn for comparison to three bands of approximately equal

[7] During our data collection, this company was known as *Crowdflower*. See Supporting Information **??** for details.

lengths—between 345–360, 360–375, and 375–390 words—to avoid comparisons where deciding on the "easier" snippet encourages coders to simply select the one noticeably shorter than the other. From this set, we randomly selected 2,000 pairs of snippets for direct comparison. To produce the estimates from the Bradley-Terry scaling, we also needed the pairs to be "connected" in these sense that every snippet must meet at least one snippet in a contest that meets others (there can be no "islands" of snippets that only meet each other).

The snippet pairs were assigned randomly to participants, in individual tasks consisting of ten comparisons each. Each pair in our dataset was rated at least twice by different coders.[8] Coders judged a median and mean of 18 and 33 pairs, respectively.

We took standard steps (e.g. Berinsky, Margolis and Sances, 2014; Benoit et al., 2016) to stop coders from generating low quality comparisons due to lack of effort, fatigue, or a skill level below the task requirements. Tasks were interspersed with "gold standard" pairs, in which one snippet is unambiguously easier than the other, at a rate of one in ten. To create the gold standard test questions, we selected some snippet pairs with the largest disparity in FRE scores, verified through inspection. Prior to being accepted for the task, a crowd worker had to pass a qualification test consisting entirely of test questions, answering at least 7 of 10 correctly. Following successful qualification, a coder performed job lots of ten pairwise comparisons, where one of these was a test question. Workers who did not maintain an overall accuracy rate of 70% correct on the test questions were removed from the pool of workers and their answers dropped from the dataset. To the 2000 pairs, and the gold standard pairs, we also added another 5% of special gold "screener" questions, designed to ensure simply that the coders were paying full attention and reading each snippet completely. These screen tasks were those whose answer from the text itself was not as obvious as with the regular gold questions, but which contained explicit, embedded instructions such as "Disregard the text and code this snippet as EASIER."

After removing duplicates, our dataset of snippets consisted of 7,236 total pairings for comparison, including 836 "gold" questions, of which 310 were screeners. We crowdsourced the com-

[8] In some very rare cases, less than one percent of all contests, a coder would see a particular pair more than once.

parisons using a minimum of three coders per pair, yielding 19,810 total comparisons, of which 13,430 did not involve screeners or test questions.

4.2 Using the pairwise data to estimate underlying textual easiness

With the human pairwise judgment data from the crowdsourcing, we were able to fit a model to estimate the latent dimension on which these texts differed, using the model for pairwise comparisons provided by Bradley and Terry (1952). This model has been applied elsewhere in political science for similar tasks (Loewen, Rubenson and Spirling, 2012; Lowe and Benoit, 2013), and we therefore give only an expedited description here following the notation of Turner and Firth (2012).

The input data is the result of our human coders having declared winners in the large number of "easiness contests" between snippets. For a given contest, crowd workers must decide which of two snippets i and j is easier to comprehend (no ties are allowed). If the easiness of these snippets are α_i and α_j, respectively, then the odds that snippet i is deemed easier than j may be written as α_i/α_j.

Defining $\lambda_i = \log \alpha_i$, the regression model can be rewritten in logit form:

$$\text{logit}[\Pr(i \text{ easier than } j)] = \lambda_i - \lambda_j. \tag{1}$$

Subject to specifying a particular snippet as a "reference snippet" (whose easiness is set to zero), this setup allows for maximum likelihood estimation of each snippet's easiness (technically, the logarithm of the easiness).

This *unstructured* Bradley-Terry model rests on several assumptions. First, that the outcomes of the contests are (statistically) independent of one another: that the result of the kth contest does not affect the result of the $k+1$th contest. Here of course, the players (the snippets) are inanimate objects so there are, for example, no "experience" effects from winning or losing. Still, it could be the case that coders see a snippet early on and deem it to have a general quality which then biases their assessment of it later (in whatever contests it appears). We are not overly concerned. For one thing, in 84 percent of the contests in our main data, the coder involved only saw a given snippet *once* in the entirety of their work for us. So any effects are likely to be small.

Second, we made no allowance for variability between snippets through any sort of random effects, either in this unconditional model or in the structured version used below, for snippets which have otherwise identical covariate values. That is, we are not using any kind of random effects for the snippets themselves. This is because we want other researchers to use our technique to model *their* data (which presumably does not contain the same exact snippets)—i.e. we care about the external portability of the work rather than the best possible local model fitting.

As a result of fitting Eq. (1) to our pairwise data, we obtained estimates of λ_i for each text snippet, as an unconditional estimate of that text's relative easiness.[9] Our task in the next step was to determine the predictors of this outcome using a separate model.

4.3 Selecting predictors using random forests

We have a large number of potential determinants of textual sophistication whose relative contribution to textual sophistication needs to be tested and fitted empirically—the 22 variables listed in Table 1. We have grouped the variables in terms of whether they refer to longer words, rarer words, longer sentences, or more difficult content. We also list the variable names associated with factor, and indicate any traditional readability measures that include these inputs. Those based on a corpus baseline of rarity, such as the Google and Brown word rarity measures, or the parts of speech and dependent clause measures, are novel to our approach.

[Table 1 about here.]

Longer words. We have various measures of word length: count of words with more than one (`W_1Sy`), two (`W2Sy`) and three syllables (`W3Sy`); count of words with less than three syllables (`Wlt3Sy`); count of words with at least six (`W6C`) or at least seven (`W7C`) letters; mean

[9] In practice, it is occasionally the case in our sample that a snippet never wins or never loses. The usual consequence of this kind of data separation would be infinite ability estimates. In one run of the model, we simply deleted those missing values, and in another we used the bias-reduction technique of Firth (1993) to ameliorate this problem. The results, in terms of the variable importance order, are essentially identical.

number of syllables per word (meanWordSyllables); and mean number of characters per word (meanWordChars).

Rarer words. We have various measures of word rarity, including membership in the Chall and Dale (1995) word list (W_wl.Dale.Chall). But for measuring more helpful *usage* rates, we drew on the frequencies of words relative to the frequency of most common word in the English language—*the*—from two large baseline corpora: the Brown corpus (Francis and Kucera, 1964) and the Google books corpus (Michel et al., 2011). For each baseline corpus, we computed a measure of its average word's relative frequency (brown_mean and google_mean) and its least frequent word's relative frequency (brown_min and google_min).

The Google book corpus offers one key advantage over the Brown corpus: it consists of unigram term frequencies specific to the year in which they were written, ranging from 1505 to 2008 (whereas the Brown corpus was collected at one point of time in the 1960s).[10] We thus obtain a relative term frequency that is specific to each year. Normalizing relative to the frequency of the term *the* provided a temporally grounded benchmark, because its relative frequency has remained relatively unchanged in several hundred years. This allowed us to compare the relative frequencies of terms without being affected by changes in overall word quantities or transcription accuracies (which vary significantly over the time in the dataset). After filtering out tokens that occurred fewer than five times or that did not match a dictionary of 133,000 English words and word forms, we ended up with a table of frequencies for 82,558 unique word types from the total Google corpus. To smooth out yearly idiosyncracies, we aggregated the term frequencies by decade.

By comparing the frequencies of SOTU addresses to their baseline frequencies in the decade they were written, we were able to distinguish between words that appear to be difficult today but were not in the decade they were written, and words that were genuinely difficult because they were rare even when written. To use our example above, the inclusion of the word *husbandry* in a contemporary speech should be considered as "harder" for a contemporary audience (such as our crowd coders) than it was in the 19th century when its use was relatively common. In Supporting

[10] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

Information A we give an intuitive example of how rarity "works" for comparing speech snippets.

Longer sentences. We have various measures of sentence length: mean sentence length (`meanSentenceLength`); mean sentence syllables (`meanSentenceSyllables`); and mean sentence characters (`meanSentenceChars`). As a final measure we divide the number of sentences by the number of characters in the snippet (`pr_sentence`).

More complex content. We measure this complexity in two ways: first, by computing the relative balance of different grammatical forms, represented by parts of speech; and second, by assessing the structure of clause dependencies using a dependency parser. It is possible that different types of words make a text more or less difficult, so for each snippet we record the proportion of nouns (`pr_noun`), verbs (`pr_verb`), adjectives (`pr_adjective`), and adverbs (`pr_adverb`). Parts of speech were identified using the spaCy NLP library (Honnibal and Montani, 2017). We give more details in Supporting Information B. To measure structural complexity, we also used spaCy to count the number of independent clauses a text contains, normalized by its length in characters (`pr_clause`).

To estimate the relationship of each variable to our Bradley-Terry estimates of a snippet's easiness, we used random forests (Breiman, 2001). This allowed us to confront two closely related estimation problems: having a large number of variables, and having very high correlations among these variables at the snippet level. Given that we want a formula that is both parsimonious and general, we need to reduce the dimensions of the problem significantly, while also having interpretable estimates. Random forests produced estimates of the relative importance of each variable to predicting the outcome, which we can then use to select the most helpful predictors. This approach's main advantages for our problem are accuracy, efficiency, an "automatic" importance measure, and a low risk of overfitting relative to other tree-based classifiers (see, Montgomery and Olivella, Forthcoming, for discussion). Other scholars, such as Montgomery and Carlson (2017), have used item response models for similar tasks. We chose random forests because we did not wish to estimate coder effects, but focused instead on producing an importance ranking of predictors. Of course other algorithms such as support vector machines allow feature ranking, but typically these require more tuning from the analyst and their task performance is not as good (see Caruana, Karampatziakis and Yessenalina, 2008, for evaluation of various methods for high dimensional data).

4.4 Using the selected predictors to fit a structured Bradley-Terry model

With the selected predictors from the previous stage, we refit the Bradley-Terry model to the pairwise contests, but in a structured form using the selected predictors as covariates. This made the easiness of the snippets conditional on a set of covariates x_{ir}, reparameterizing the easiness λ_i of a given snippet as

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir}. \tag{2}$$

From this structured Bradley-Terry model, the estimated $\hat{\beta}$ coefficients tell us the marginal effect of each x-variable on the perceived (relative) easiness of the snippets. Once the $\hat{\beta}$s are obtained, we can used these to predict a $\hat{\lambda}_i$ for any (new) text for which we can compute the necessary covariate values.

In fitting the structured model, we did not include so-called "contest-specific predictors" either indirectly—such as effects for (the proclivities of) given human coders—or directly by allowing for consequences of the order in which the snippets were presented to the subjects who judged them. Not only are such effects hard to estimate in a world in which the median coder only performs 18 comparisons, but also experience shows that once filtered through the minimum quality threshold, no relevant coder characteristics remain that correlate with coding decisions in a way that materially affects the quality of the resulting data (see Benoit et al., 2016, for discussion).

5 Results

With the unstructured Bradley-Terry estimates of each snippet's textual "easiness", we were then in a position to determine which of our potential set of 22 predictors best predicts easiness, and to what extent, in our context and to compare this with more traditional measures. Prior to this, we compared the traditional approaches with one another as described in Supporting Information D: they do essentially equally well, but we focus on comparisons with FRE, because it is most familiar, in what follows.

5.1 Fitting the structured model to the training data

To gauge the contribution of our candidate predictors on the unstructured "easiness" measures, we compared the random forests results in terms of model fit and percent correctly predicted.[11]

Our initial supervised model suggested the key predictors of easiness were the time-specific rarity of the least frequently used word (`google_min`), the average sentence length measured in characters (`meanSentenceChars`), and the proportion of nouns (`pr_noun`). These variables collectively were both small in number (allowing for a simple formula) and most "important" in the random forests specific sense discussed in Supporting Information E. In relation to the classical measures, sentence length has long been a common element of readability indexes, but our inclusion of word rarity and the measure of nouns is novel.

To assess the performance of this model in predicting the pairwise contests, and to compare it to the most common classical measures, we constructed a baseline model which uses the FRE as its (only) covariate content. We did this in two ways. First, we include the FRE of the snippet using the weights from Flesch's (1948) original formula. Second, we include the variables Flesch (1948) includes, but allow the model to calculate the optimal weights for our political data. In Table 2 we report the findings from those models, in the leftmost two columns. For the "FRE baseline" model (original weights) we see that the Akaike information criterion (AIC) is 26267.79, while the augmented proportion (of contests in the data) correctly predicted (PCP) is 0.719. When we allow the weights on the relevant variables to adjust to local conditions (column 2) we see a commensurately better model fit: the AIC falls to 25910.29, and the proportion correctly predicted rises to 0.737. This is in line with our thinking above: in particular, that models work best when fit

[11] Interpreting our results requires dealing with a subtle problem in calculating the denominator of the model proportion correctly predicted, that stems from the fact that the coders do not always agree on a "correct'" answer. We adjust for this in a way described in Supporting Information C.

to relevant data. Column 3 represents our basic three variable model as discussed above. Clearly, it does better than the Flesch model with the original weights, but—perhaps surprisingly—not as well as the re-weighted version (AIC is higher).

[Table 2 about here.]

This model ("Basic RF Model") did not include a measure of word length, despite this feature being one of the two core components of the FRE. From the results of variable importance (presented graphically in Supporting Information E), the most important measure of word length that predicts easiness is the average number of characters per word (`MeanWordChars`). We added this variable to our machine learning model and re-fit the structured Bradley-Terry model, shown in the fourth column of Table 2 (headed "Best Model"). This model outperforms every other version, with the lowest AIC (25740.25) and the highest PCP (0.741). In an effort to ascertain the robustness of this model, we dropped the parts-of-speech variable (`pr_noun`) and added the next highest rated one (`pr_adjective`): the fit of the model was essentially identical. In what follows we work with the one that uses the part of speech—nouns in this case—that the learner preferred in importance terms.

While full details of the random forest models that we ran on the unstructured abilities, along with variable importance plots, are provided only in Supporting Information E, we note that all variable effects were both in the expected directions and statistically significant at conventional levels. The higher the relative frequency of the least frequent word (relative to *the*), the easier the snippet was to understand. Snippets that contain longer sentences and longer words were both judged to be less easy to understand. Finally, increasing the proportion of nouns was also associated with increased easiness.

To gauge the significance of the differences in accuracy of the reported models, we provide a bootstrapped 95% confidence interval on the percent correctly predicted, based on 500 sentence-level resamples. Our key observation is that the confidence interval for the fit of the final model does not overlap with the FRE model, implying that it is indeed better in a statistical sense. On what types of data, exactly, does our model perform better? Unsurprisingly, it performs best when two documents are similar other than the proportion of nouns they contain, or the rarity of their words. In the contests for which our model outperforms the Flesch version to the greatest extent, it is the word rarity input that matters most. To get a sense of this, compare these two snippets. The first is from Obama's 2009 address, and has an FRE of around 50:

> I speak to you not just as a President, but as a father, when I say that responsibility for our children's education must begin at home.

The second is from Cleveland's 1889 effort,[12] which has an FRE of approximately 67:

> The first cession was made by the State of New York, and the largest, which in area exceeded all the others, by the State of Virginia.

The FRE model predicts this to be a relatively straightforward win for Cleveland's speech. Our model, of course, penalizes the estimate of its simplicity due to the presence of the relatively rare term *cession* (along with there being slightly fewer nouns in the second document). Indeed, the frequency of the least common term in Obama's speech is over three orders of magnitude larger than that of Cleveland's speech—something that our approach clearly captures but that traditional indices cannot.

It is helpful to be candid about several issues pertaining to our results. First, clearly, while we are outperforming the most widely-used measure of readability, our gains are not huge in an *absolute* sense. The largest gains in predictive accuracy come from refitting the Flesch model appropriately to the data rather than using its usual "off-the-shelf" mode. Nonetheless, these gains are reasonable in a *relative* sense. The baseline Flesch predictive accuracy was 71.9%—about 22 percentage point better than chance. Our final model is 24.3 percentage points better than chance, a relative increase of around 11%. But this increase is "real" per our discussion of the bootstrap results above. Third, whether or not one uses our *specification*, the general approach—of training on relevant data and providing model-based estimates—is preferable for the reasons given above. Even if one simply wanted to use the Flesch set up (in terms of its component variables) based on Table 2 we would recommend fitting to domain data for that purpose.

[12] This snippet appears per our discussion in Supporting Information **??** about including some older texts from an earlier pilot study.

5.2 Applying probabilistic comparisons

Using the fitted four-covariate "Best Model" from Table 2, we can estimate a fitted easiness score for any text. There are two ways of applying this model. First, given Equations (1) and (2), we can obtain a (point) estimate of the probability that any given text i is easier (or conversely, more difficult) than any other text j by calculating

$$\Pr(i \text{ easier than } j) = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)}. \tag{3}$$

To see how this works, consider two snippets, one from Clinton (1999),

> If we do these things—end social promotion; turn around failing schools; build modern ones; support qualified teachers; promote innovation, competition and discipline-then we will begin to meet our generation's historic responsibility to create 21st century schools. Now, we also have to do more to support the millions of parents who give their all every day at home and at work.

and one from George W. Bush (2005)

> And the victory of freedom in Iraq will strengthen a new ally in the war on terror, inspire democratic reformers from Damascus to Tehran, bring more hope and progress to a troubled region, and thereby lift a terrible threat from the lives of our children and grandchildren. We will succeed because the Iraqi people value their own liberty—as they showed the world last Sunday.

[Table 3 about here.]

For each of these snippets, Table 3 gives the relevant covariate values for our best model above. Using the coefficients from Table 2, it is a simple matter of matrix multiplication to form the $\hat{\lambda}$ values and to compute the probability that the Clinton text is easier than the Bush text.[13]

$$\begin{aligned}
\Pr(\text{Clinton snippet easier than Bush snippet}) &= \frac{e^{\lambda_{\text{Clinton}}}}{e^{\lambda_{\text{Clinton}}} + e^{\lambda_{\text{Bush}}}} \\
&= \frac{\exp(-2.64)}{\exp(-2.64) + \exp(-2.92)} \\
&= 0.57.
\end{aligned}$$

[13] Just for presentational sanity here, we are rounding all values. This has inevitable precision loss, and values produced by our software will differ in practice for such examples.

We can also compare each text to a common baseline text, for instance to a corpus of texts judged to be at a fifth-grade reading level. We obtained examples of such texts from a university education department,[14] and estimated the relevant λ to be -2.184507. Thus, the probability that the Clinton text is easier than a fifth grade text is estimated to be 0.259, and the probability that the Bush text is easier to follow than the fifth grade works is 0.209.[15] We can place confidence intervals around the point prediction by bootstrapping the sentences in the texts (in the sense of Lowe and Benoit, 2013), where each replicate produces a new computation of the covariate values and then used to compute fitted values given the estimated model. Note that the differences between texts mean something extremely well-defined here: we can make concrete statements about *how much* easier one document is relative to another, and the quantity refers back to a sensible model. This is quite unlike FRE, for which a difference of 5 points on the scale has no natural, cardinal interpretation.

Along with model-based estimates, researchers may also want a quantity analogous to the continuous 0–100 scores from the Flesch (1948) (regression) formula. In Figure 1, we have rescaled the λs (that is, the $\mathbf{X}\beta$s, without applying the exponential function) such that texts measured to be at the fifth grade level receive a value of 100 and those at the post-college level a value of 0.[16]

[Figure 1 about here.]

Experimenting with the continuous measure on the SOTU snippet corpus performs well in the sense that it returns point estimates on a roughly 0–100 scale commensurate (but not identical) to the FRE equivalents. This works because it replaces a logit-style calculation that is not linear in the predictors with a linear sum (i.e $\sum_{r=1}^{p} \beta_r x_{ir}$), exactly like the regression-based formula for FRE. In Figure 1 we provide a scatterplot of our measure for the snippets (y-axis) relative to the FRE for the same data (x-axis), along with the line of linear fit. The correlation over the full range of points (~ 0.7) is reasonably large and positive. Within the (theoretical) minimum and maximum of the

[14] `https://projects.ncsu.edu/project/lancet/fifth.htm`

[15] Here we are using computer precision for our calculations.

[16] We used the collection of fifth grade texts we mentioned above for the easy end of the scale, and the most difficult snippet (which had an FRE of around 3) for the “hard” end.

FRE range of 0-100, however, the correspondence is even higher. This implies that for the great majority of documents for which FRE is used, our measure—preferred on theoretical grounds—is a good choice that will behave as expected. Outside the 0-100 range, particularly to the bottom left of the plot, our measure tends assign a considerably harder score for the hardest texts.

6 Reanalyzing the State of the Union addresses

Just as we demonstrated how to apply the fitted model to other short texts to estimate their easiness level, we can also apply this to each SOTU address in its entirety. Using our model-based probability measure—here, with a fifth grade text as a baseline for comparison—Figure 2 plots the relevant point estimates and 95% (simulated) confidence intervals (*y*-axis) plotted against the date of the relevant text. The probability estimates are drifting upwards over time, but generally stay below 0.50. But because we are using a well-defined statistical model, we can say more about the data. In particular, the confidence intervals allow us to make comments about sampling uncertainty. Note that there is considerable overlap between the intervals for the post-war period (for example, some of the speeches in the early 2000s are not so different to those in the early 1950s). This implies that statements about the simplification of language may be correct in some aggregate sense if we consider the entire period since the founding of the Republic, but less clear for modern times specifically.

[Figure 2 about here.]

For the closest equivalent to a direct comparison with more traditional approaches, Figure 2 plots the ratio of the FRE for each SOTU speech compared to a corpus of texts designated to be at the fifth grade reading level,[17] and shown by the smoothed loess line. The measures agree in terms of general direction—addresses become easier over time—but differ in terms of magnitude. In particular, our measure has the speeches prior to around 1910 being considerably more difficult

[17] This is divided by two to ensure normalization in the sense that two texts of equal difficultly should have probability 0.50 of beating each other.

to understand than FRE claims they were. Post-1910, our measure tends to have the estimated ease of understanding the passages as higher than FRE. To the extent that one believes that new technology, such as the radio and the television, lead to speeches that are easier to follow after the first decade of the 20th Century, this makes sense. And, to reiterate, our model is actually trained on appropriate, political data with local, decade specific, word rarity measures.[18]

7 Comparing on a dimension of political interest

A pleasing feature of our approach is that it facilitates direct comparison of texts that differ with respect to some meta-data or covariates of interest to produce probabilistic statements about those differences. To demonstrate this, we compare the complexity of *written* SOTU addresses differ to that of *spoken* ones. Because the former medium of delivery was historically much more prevalent and the latter is the norm now, meaningful comparisons are difficult because there is obvious confounding over time and across authors. In 1945, 1956, 1972, 1974 and from 1978-1980, however, each president delivered two SOTU speeches, one spoken and one delivered in writing to Congress, on the same day, and on the same topics. Since all else was generally equal except the medium of communication, this allows us to compare directly the degree of textual sophistication for written versus spoken texts.

[Figure 3 about here.]

Figure 3 plots the results, showing probability that the spoken address was easier than its written counterpart. Across the set of seven paired addresses, the probability was between about 0.54 and 0.64 that the spoken address was easier. Speculatively, this may help to explain recent trends to easier and easier addresses by presidents: they are giving them as speeches rather than as written text. Our framework makes this comparison possible using explicit probability statements.

[18] In Supporting Information F we look at the way that our "dynamic" adjustments affect our aggregate estimates: the differences are not huge, but are in the expected direction.

8 Summary and Discussion

The nature of the messages that political actors send one another are of key interest to political science, whether it be in American politics, international relations or from a comparative perspective. Yet a curious gulf has emerged in our studies. On the one hand, we have plenty of theory and empirical evidence that such communication matters: whether it be "dog whistle" in nature (Albertson, 2015), rhetorical (Riker, 1996), vague (Lo, Proksch and Slapin, 2016), or more explicitly designed to appeal to certain types of agents. On the other hand, the discipline has been slow to adopt textual complexity measures in any context. This is despite the fact that the various readability measures are easy to use and scale in a straightforward way, which is important given the sheer amount of textual data now available to scholars. Presumably, part of this reticence is lack of familiarity with such approaches. But part of it is likely a very reasonable skepticism about the merits of these educational measures—a concern echoed in other fields of social science (e.g. Sirico, 2008; Loughran and McDonald, 2014) and indeed, increasingly in education itself (Ardoin et al., 2005).

Rather than attempt to rehabilitate the indices, here we focused on producing something better, considering all possibly relevant inputs, using a statistical method for determining which inputs explain textual complexity in our context and how, using an explicitly comparative framework built on pairwise comparisons. In Table 4 we summarize our contribution relative to problems in using traditional readability measures to estimate the textual sophistication of political text.

[Table 4 about here.]

To get the pairwise comparisons needed to fuel our context-based estimates, we used human coders (via the crowd) to provide relative assessments of short texts, and from there we built a well-defined statistical model. That model uses variables that differ from standard approaches, including word rarity and parts-of-speech information. The final version performs better in fit terms too, although precisely because the approach is on much firmer probabilistic grounds it is hard to compare directly to previous approaches. Fundamentally then, we have improved practice here: the approach is transparent, sensible and model-based and trained on relevant domain data. It is also flexible, in the sense that the workflow and software we have designed allows end-users to calibrate the method to their specific problems.

On the question raised in our introduction—"is discourse being dumbed down?"—our purpose here is less to provide a decisive answer as to provide tools for more accurately answering this question. Certainly, the State of the Union addresses have become easier to comprehend in the modern era. The actual political *sophistication* of a political message, however, depends more on the content of the message. Traditional measures based on static indexes of readability are unable to capture this directly: for example, shorter sentences may be a good or bad thing, depending on the context. And that context is more likely to be captured via local fitting (to the type of text at hand), measuring the grammatical structure of the documents, and the rarity of the terms they use. These are precisely the things our approach *can* model. By outlining a flexible approach to the problem, furthermore, we facilitate comparisons of different inputs' effect on textual sophistication, allowing more precise answers to the sources and nature of the trend to growing sophistication in political speech. Finally, we note that prior to our efforts here that use historical benchmarks for familiarity, we had very little idea about whether the documents in question were unusual relative to what readers would have experienced at the time. That is, although not the focus of our work here, we can get some sense of how similar the structure of, say, the SOTU of 1815 was to other readings on offer that year. Put crudely, if the SOTU from that time is much more erudite than the one in 2015, but simultaneously much harder to understand than the *average* (Google books) text in 1815, it gives claims about dumbing down a very different complexion.

While our contribution will be helpful for those interested in characterizing political communication, it is hardly the last word on the matter. We have provided a statistical machinery, and variables, for thinking more carefully about the measurement of sophistication or clarity in texts. What we have not done is produced a straightforward way to distinguish between more subtle understandings of such concepts. For example, one can imagine a politician—a president of the United States even—who uses relatively common terms in simple sentence constructions, but is not especially clear. By contrast, great academic writers might be able to describe extremely complicated ideas in straightforward ways for popular audiences. Our approach would generally be better than previous ones, but is still unlikely to place these two extremes correctly on the same scale. This is, of course, because a sophisticated idea (like democracy, or inclusivity or conservatism) need not be complicated in expression, and vice versa. More attempts should be made—not least at the coding/crowdsourcing level—to iron out these differences, possibly by introducing different dimensions of complexity at the point of testing or modeling. A related next step would be to make all of the variables dynamic, for instance, measuring the proportion of nouns in a text relative to a baseline noun usage from the time the document was written. We leave such efforts for future work.

References

Albertson, Bethany. 2015. "Dog-Whistle Politics: Multivocal Communication and Religious Appeals." *Political Behavior* 37(1):3–26.

Ardoin, Scott P, Shannon M Suldo, Joseph Witt, Seth Aldrich and Erin McDonald. 2005. "Accuracy of Readability Estimates' Predictions of CBM Performance." *School Psychology Quarterly* 20(1):1–22.

Benoit, Kenneth, Drew Conway, Benjamin Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–295.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

Bischof, Daniel and Roman Senninger. 2018. "Simple politics for the people? Complexity in campaign messages and political knowledge." *European Journal of Political Research* 57:473–495.

Bonsall, Samuel B, Andrew J Leone, Brian P Miller and Kristina Rennekamp. 2017. "A plain English measure of financial reporting readability." *Journal of Accounting and Economics* 63(2-3):329–357.

Bradley, Ralph and Milton Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39(3/4):324–345.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Bruce, Bertram, Andee Rubin and Kathleen Starr. 1981. "Why readability formulas fail." *IEEE Transactions on Professional Communication* 24(1):50–52.

Cann, Damon, Greg Goelzhauser and Kaylee Johnson. 2014. "Analyzing Text Complexity in Political Science Research." *PS: Political Science & Politics* 47:663–666.

Caruana, Rich, Nikos Karampatziakis and Ainur Yessenalina. 2008. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*. ACM pp. 96–103.

Chall, Jeanne Sternlicht and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Charteris-Black, Jonathan. 2011. *Politicians and rhetoric: The persuasive power of metaphor*. Springer.

Coleman, M and T Liau. 1975. "A computer readability formula designed for machine scoring." *Journal of Applied Psychology* 60(2):283–284.

Dale, Edgar and Jeanne Chall. 1948. "A Formula for Predicting Readability." *Educational Research Bulletin* 27(1):11–20.

Diamond, Larry. 2002. "What Political Science Owes the World." *PS: Political Science & Politics Online Forum* pp. 113–27.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27–38.

Flesch, Rudolf. 1949. *The Art of Readable Writing*. New York: Harper.

Flesch, Rudolph. 1948. "A new readability yardstick." *Journal of Applied Psychology* 32(3):221–233.

Francis, W. N. and H. Kucera. 1964. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Providence, Rhode Island: Brown University.

Fry, Edward. 1968. "A Readability Formula That Saves Time." *Journal of Reading* 11(7):513–578.

Gatto, John Taylor. 2002. *Dumbing us down: The hidden curriculum of compulsory schooling*. Vancouver: New Society Publishers.

Gibson, Edward. 1998. "Linguistic complexity: locality of syntactic dependencies." *Cognition* 68(1):1–76.

Gunning, Robert. 1952. *The Technique of Clear Writing*. New York: McGraw-Hill.

Honnibal, Matthew and Ines Montani. 2017. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." *To appear* .

Jansen, David-Jan. 2011. "Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies." *Contemporary Economic Policy* 29(4):494–509.

Kincaid, J Peter, Robert Fishburne, Richard Rogers and Brad Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy Enlisted Personnel*. Vol. Research Branch Report 8-75 Naval Air Station Memphis: Chief of Naval Technical Training.

Kintsch, Walter and Douglas Vipond. 1979. Reading Comprehension and Readability in Educational Practice and Psychological Theory. In *Perspectives on Memory Research*, ed. L. Nilsson. Psychology Press pp. 329–365.

Klare, George. 1963. *The measurement of readability*. Ames, Iowa: University of Iowa Press.

Kristof, Nicholas. 2014. "Professors, We Need You!" `https://www.nytimes.com/2014/02/16/opinion/sunday/kristof-professors-we-need-you.html`. New York Times, Online, February 15, 2014.

Lim, Elvin. 2008. *The Anti-Intellectual Presidency*. New York: Oxford University Press.

Lo, James, Sven-Oliver Proksch and Jonathan B Slapin. 2016. "Ideological clarity in multiparty competition: A new measure and test using election manifestos." *British Journal of Political Science* 46(3):591–610.

Loewen, Peter, Daniel Rubenson and Arthur Spirling. 2012. "Testing the power of arguments in referendums: A Bradley–Terry approach." *Electoral Studies* 31(212-221).

Loughran, Tim and Bill McDonald. 2014. "Measuring Readability in Financial Disclosures." *The Journal of Finance* 69(4):1643–1671.

Lowe, Will and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

Luskin, Robert C. 1990. "Explaining political sophistication." *Political Behavior* 12(4):331–361.

michalke, m.eik. 2017. *koRpus: An R Package for Text Analysis*. (Version 0.10.2).
URL: *https://reaktanz.de/?c=hacking&s=koRpus*

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant et al. 2011. "Quantitative analysis of culture using millions of digitized books." *Science* 331(6014):176–182.

Montgomery, Jacob and David Carlson. 2017. "A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts." *American Political Science Review* 111(4):835–843.

Montgomery, Jacob and Santiago Olivella. Forthcoming. "Tree-based models for political science data." *American Journal of Political Science* 62(3):729–744.

Owens, Ryan and Justin Wedeking. 2011. "Justices and Legal Clarity: Analyzing the Complexity of Supreme Court Opinions." *Law & Society Review* 45(4):1027–1061.

Riker, William H. 1996. *The strategy of rhetoric: Campaigning for the American Constitution*. New Haven: Yale University Press.

Sherman, Lucius. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn.

Sirico, Louis J. Jr. 2008. "Readability Studies: how technocentrism can compromise research and legal determinations." *Working Paper Series* 104.
URL: *http://digitalcommons.law.villanova.edu/wps/art104*

Smith, Frank. 1986. *Understanding Reading: A Psycholinguistic Analysis of Reading AndLearning to Read.* Lawrence Erlbaum Associates, Incorporated.

Spache, George. 1953. "A new readability formula for primary-grade reading materials." *The Elementary School Journal* 53(7):410–413.

Spirling, Arthur. 2016. "Democratization of Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *Journal of Politics* 78(1):120–136.

Spriggs, James F. II. 1996. "The Supreme Court and Federal Administrative Agencies: A Resource-Based Theory and Analysis of Judicial Impact." *American Journal of Political Science* 40:1122–1151.

Thurstone, L. L. 1927. "A law of comparative judgment." *Psychological Review* 34(4):273–286.

Turner, Heather and David Firth. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package." *Journal of Statistical Software* 48(1):1–21.

Wheeler, Lester and Edwin Smith. 1954. "A practical readability formula for the classroom teacher in the primary grades." *Elementary English* 31:397–399.

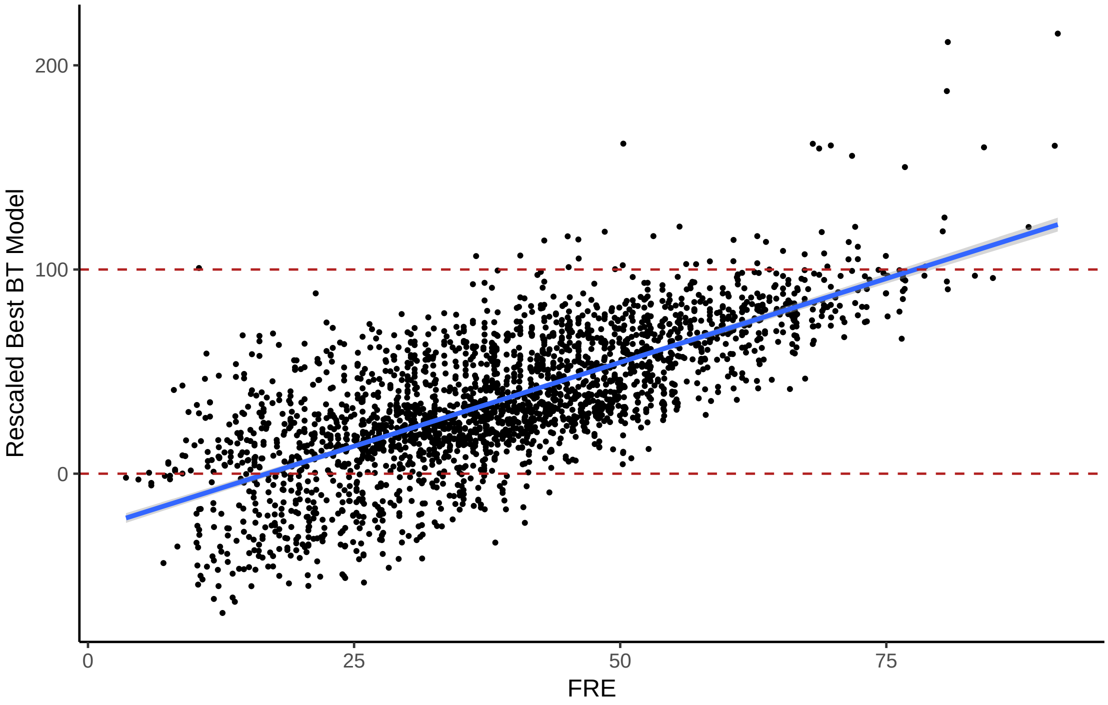

Figure 1: *Comparing our rescaled measure to the FRE of the snippets.*

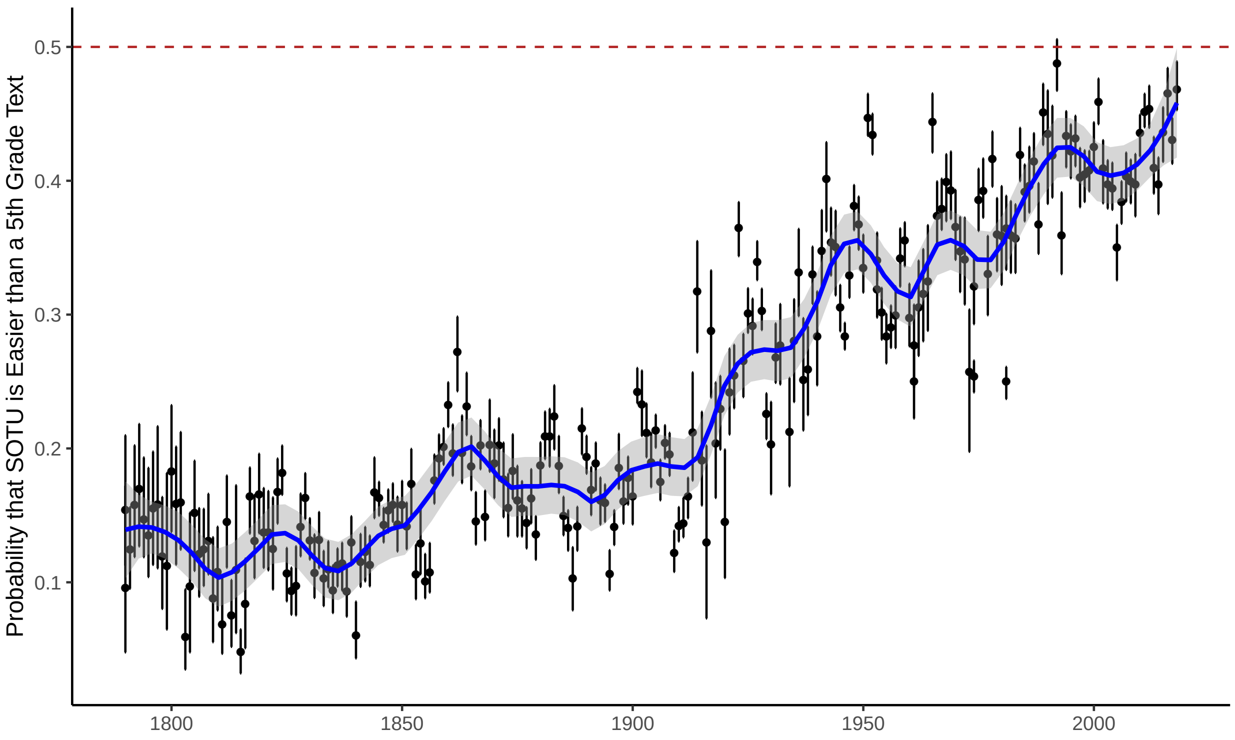

Figure 2: *Probability that a State of the Union address is easier to understand than a fifth grade text baseline, compared to FRE.* The points and associated vertical lines are probability estimates and 95% confidence intervals for our measure. The blue line is the loess fit of half the ratio of the FRE for the SOTU to the FRE of the 5th grade text corpus.

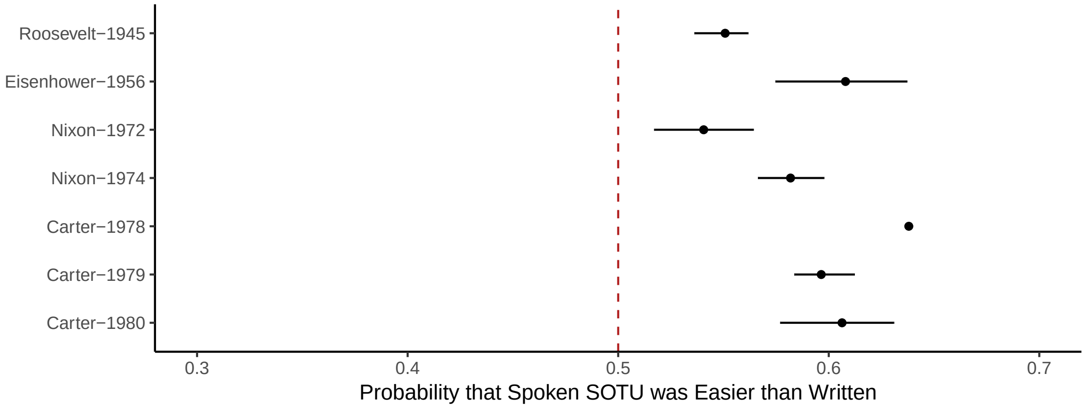

Figure 3: *Probability that a spoken SOTU address was easier to understand than its written counterpart.* The lines represent the 95% confidence intervals from bootstrapping.

Table 1: *Determinants of textual complexity.*

Source of complexity	Variable name	Used by these measures
Long words		
Mean characters per word	meanWordChars	ARI; Bormuth; Coleman-Liau
Words with at least 7 characters	W7C	LIX
Words with at least 6 characters	W6C	Harrison-Jacobson
Mean syllables per word	meanWordSyllables	Flesch; Flesch-Kincaid; Tuldava
Words with at least 3 syllables	W3Sy	FOG; SMOG
Words with fewer than 3 syllables	Wlt3Sy	FOG-NRI
Words with 2 syllables	W2Sy	ELF; Wheeler-Smith
Words with 1 syllable	W_1Sy	FJP; FORCAST
Rare words		
Google Books baseline usage	google_min	(new)
	google_mean	(new)
Brown corpus baseline usage	brown_mean	(new)
	brown_min	(new)
Words in the Dale-Chall list	W_wl.Dale.Chall	Dale-Chall; Bormuth; Spache
Long Sentences		
Mean characters per sentence	meanSentenceChars	Danielson-Bryan
Mean sentence length in words	meanSentenceLength	Flesch; Flesch-Kincaid; ARI; Bormuth Dale-Chall; FJP; FOG; Spache; LIX; Tuldava; Wheeler-Smith; Harrison-Jacobson
Number of sentences per character	pr_sentence	Coleman-Liau
Mean sentence length in syllables	meanSentenceSyllables	Strain
Complex Content		
Proportion of nouns	pr_noun	(new)
Proportion of verbs	pr_verb	(new)
Proportion of adjectives	pr_adjective	(new)
Proportion of adverbs	pr_adverb	(new)
Average subordinate clauses	pr_clause	(new)

Summary of existing measures taken from michalke (2017)

Table 2: *Comparing the performance of the structured models.* Standard errors in parentheses.

	FRE Baseline	FRE re-weight	Basic RF model	Best Model
FRE	0.02			
	(0.00)			
mean Sentence Length		-0.06		
		(0.00)		
mean Word Syllables		-1.79		
		(0.07)		
Minimum Google books rarity			1298.14	1318.65
			(153.07)	(155.64)
mean Sentence Chars			-0.01	-0.01
			(0.00)	(0.00)
noun proportion			0.43	0.31
			(0.17)	(0.17)
mean Word Chars				-0.31
				(0.02)
N	19430	19430	19430	19430
AIC	26267.79	25910.29	25915.01	25740.25
Prop correctly predicted	0.719	0.737	0.738	0.741
[95% CI]	[0.710, 0.727]	[0.728, 0.747]	[0.729, 0.748]	[0.733, 0.751]

Table 3: *Examples of covariates from two snippets in the data.*

Variable	Clinton	Bush
Min Google rarity when speech given	2.65e-04	1.40e-08
Mean Sent Chars	155.50	153.50
noun proportion	0.30	0.23
mean Word Chars	4.94	4.72
$\hat{\lambda}_i$	-2.64	-2.92

Table 4: *Summary of our approach as a solution to a series of problems with traditional approaches.*

Dimension	*Traditional Approach*	*Our Approach*
Development context	Education research	Political text
Test subjects	Schoolchildren	Adults
Temporal context	Readers in 1940s/50s, not easily updated	Contemporary readers, easy to update (via crowdsourcing)
Assessing model fit	Cannot assess quality/fit of predictions for documents	Straightforward to assess absolute model fit (in training set) via usual metrics like percent correctly predicted
Comparison of different measures	Cannot compare models of different forms	Straightforward to assess relative model fit (in training set) via usual metrics like AIC, BIC
Interpreting differences	Cannot interpret fine-grained differences in document scores	Natural model-based interpretation of document estimates (via Bradley-Terry model)
Uncertainty accounting	No uncertainty around estimates.	Uncertainty estimates available both for variables in model, and on document scores (via bootstrapping)
Selecting inputs and assigning weights	Composite indices/aggregate form hides changes in input variables	Straightforward to examine all changes to component parts
Rarity of term usage	Rarity of terms accounted for in *ad hoc* inflexible way, if at all	Rarity of terms systematically derived from large corpus, and available for any period of interest in past 200 years.