1 Methods: Crowdsourcing Complexity

With the above considerations in mind, we aim to "discover" the textual features that constitute complexity. We employ human coders in the crowd to perform hundreds of pairwise comparisons to evaluate which of a pair of snippets is more complex. We drew these snippets from 70 US State of the Union (SOTU) Addresses, dating from 1950. This corpus is appropriate because, while it spans hundreds of years and the concomitant variations in linguistic styles, its purpose has remained relatively constant. We restricted our use to relatively recent speeches because the lexicon of the postwar era is relatively comparable, and because from the early twentieth century onward, many speeches were delivered in written form. Traditional readability scores like FRE have also been applied to SOTUs in the popular press[1], and changes in this measure have substantive implications for the nature of accountability and democracy in the United States.

1.1 Preparing the Snippets: Gold Standards and Crowdflower

We begin with the raw texts of the 70 SOTU Addresses. Until about 1913, written copies of these Addresses were delivered to the Congress, while afterwards they were presented directly to the public via radio and later television. Each form of address brings some organizational non-sentence pieces of text, which we remove. We broke the Addresses up into one- and two-sentence snippets of text.[2][**KEN: verify para**]

These snippets vary dramatically in the number words they contain. This is clearly an important component of textual complexity–so important, in fact, that we take it as given and measure the variation in complexity among snippets of similar lengths. We match snippets of approximately equal length, to avoid comparisons where deciding on the "easier" snippet appears easy because one is noticeably shorter than the other. Along the same lines, we take the FRE scores as the baseline against which our measure should be compared. Within each group of snippets of similar

[1] *The Guardian*, Feb 12, 2013., https://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level.

[2] We remove any snippets containing large numbers, as these often appeared as part of a list or other non-natural language construction.

lengths, we sort the snippets once by FRE scores in ascending order and again in descending order, and combine these two lists to create a list of comparisons that vary from extremely disimlar to extremely similar FRE scores. [**KEN: verify para**]

Similar to Benoit et al. (2016), we then use the Crowdflower platform to recruit coders to evaluate which of each snippet is more complex. We labeled the task as "Identify Which Of Two Text Segments Contains Easier Language." Upon accepting the task, we provide the workers with a number of example comparisons, with one option correctly labeled as more complex. For a visual representation of the interface, see Appendix A. The specific instructions provided to each worker follow:

> 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.

A crucial aspect of crowdsourcing any coding task is ensuring that workers provide high quality responses. To that end, we employ gold standard tasks: interspersed among the comparisons of interest are comparisons in which one snippet is unambiguously more complex than the other, at a rate of one test question in ten. To create the gold standard questions, we select the 10% of the snippet pairs with the largest disparity in FRE scores, verified through inspection. If the worker incorrectly classified more than 30% number of these gold standard tasks, she was removed from the pool of workers and her answers removed from the dataset.[3] Prior to being accepted for the task, a crowd worker also had to pass a qualification test consistently entirely of test questions, answering at least 7 of 10 correctly.

[**KEN: verify para**] To create the snippets, we formed all one- and two-sentence segments from the State of the Union corpus, ranging between 100–300 and 180–400 characters respectively. After discarding snippets whose Flesch scores were lower than 10 or higher than 100, we

[3] Following Berinsky, Margolis and Sances (2014), we also included some "screener" questions, which appear to be the same as normal comparisons but include at some point the phrase "Disregard the content and code this sentence as EASIER." Appoximately 10% of the test questions were screeners, and approximately 10% of the total comparison tasks were test questions.

were left with 43,358 snippets, of which about two-thirds were the single-sentence texts. From these, we randomly sampled 2,000, and from these formed 10,000 pairs in a way that guaranteed the connectivity of pairs for comparison to enable Bradley-Terry scaling. For the results shown below, we crowd-sourced the comparisons of 20% of these, specifying a minimum of three coders per pair, yielding 6,000 data points (excluding 1,997 test comparisons also coded) on a total of 1,977 unique snippets.[4] To aid the automation of this process and to provide both reproducibility and transparency, we implemented all of the functions to sample snippets, create pairs and test questions, prepare the data for Crowdflower, and to process the crowd-coded data in an R package sophistication, which also includes the cleaned version of the SOTU corpus.

1.2 Incorporating Familiarity: Google n-grams and parts of speech

[**KEN: verify para**] Because our test data spans political speech dating to the 1790s, we also took a corpus linguistics apprach to benchmark word usage as a measure of how uncommon (and hence how difficult to understand) were the words in a text. For this purpose, we incorporated a benchmark of how unusual words were in contemporary context. To this end, we downloaded the unigram frequency datasets from the Google Book corpus,[5] which yields token counts on a yearly basis from 1505 until 2008. Processing this enormous data and discarding any years prior to 1790 resulted in a total set of 615,362,456,717 token counts from 85,623 word types, 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.[6] To smooth out individual differences in the yearly samples, we combined the frequency counts by decade.

[**KEN: verify para**] To assess the benchmark of a how unusual was a text, we computed the frequency of each term in a decade relative to the frequency of the word "the", the most common word in the English language and also one whose relative frequency is relatively unchanged in

[4] The difference from the 2,000 pairs chosen as to do with the fact that we ran only a 20% sample of the total job of 10,000 pairs.

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

[6] One reason for the massive drop in the number of word types is that many appear to be artifacts of errors introduced in optical character recognition. (You should have paid more attention to typing those CAPTCHAs correctly.)

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 sampled). For instance the word "husbandry" (the cultivation and breeding of crops and animals) was used 8.5 times more frequently in the 1790s than it was in the 2000s. Because we think this would make it harder for a contemporary audience (such as our crowd coders), we computed the score of each text based on the frequencies from the decade of the 2000s. Our computation looked up every token in a text from the Google unigram frequency corpus in 2000, and computed a mean relative word frequency as well as the relative frequency of the *least* frequent word.

For example, considering the following two snippets:

1. "Numerous are the providential blessings which demand our grateful acknowledgments. . . too important to escape recollection." (George Washington, 1791)

2. "Now, we have to build a fence. And it's got to be a beauty." (Donald Trump, 2015)

These are 15 and 14 tokens in length, but the mean frequency relative to *the* in the 2000s for the first was 0.11, and 0.14 for the second, indicating that the mean word in Washington's speech was relatively much less frequently used than in Trump's. The word that is used least commonly (relative to *the*) in the two snippets induces a large difference in the measurements of the texts: for Washington, it is *providential* which has a ratio of 0.00002085 relative to *the* (implying *the* is used about $48,000$ times as often). For Trump, the relevant word is *fence*, for which the ratio is an order of magnitude higher, at 0.00025 (meaning *the* is used about 4000 times as often). (We note also that the Flesch Reading Ease for the Washington text is 5.5, compared to 105.1 for the Trump snippet.)

[**KEN: verify para – add new package info**]We also computed the relative frequency of parts of speech in each text, tagging the snippets using the Google Universal tagset[7] using the Python-based spaCy toolset.[8] This follows some readability indexes, such as Tränkle and Bailer (1984),

[7] See https://github.com/slavpetrov/universal-pos-tags.

[8] See http://spacy.io.

that consider conjunctions and prepositions, and Coleman's "C3" and "C4" indexes (Coleman and Liau, 1975) that take into account the frequency of pronouns and propostions. Converting these to relative frequencies for each snippet gave us information on the proportions of nouns, adjectives, verbs, prepositions, and so on. This allowed us to include these quantities in the feature set for fitting models below to predict reading ease.

1.3 Bradley-Terry Regression Analysis

Exposition of the Bradley-Terry model (Bradley and Terry, 1952) can be found in numerous textbooks (e.g. McCullagh and Nelder, 1989), but we follow the presentation found in Turner and Firth (2012) for our work here. The input data is the result of our human coders having declared winners in the large number of "contests" between snippets. For a given contest, crowd workers must decide which of two snippets i and j for is 'easier' to comprehend (no ties are allowed). If the 'easiness' of i is α_i, and the 'easiness' of j is α_j, then the odds that snippet i is deemed easier than j may be written as $\frac{\alpha_i}{\alpha_j}$.

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 set-up allows for maximum likelihood estimation of each snippet's easiness. For current purposes though, we wish to make the easiness of the snippets a product of covariates—that is, the average length of words they contain, the number of syllables the words have etc. This is achieved by modeling the easiness of a given snippet as

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

This is known as the *structured* Bradley-Terry model: the set of β coefficients then tells us the marginal effect of each x-variable on the perceived (relative) easiness of the snippets. Notice further that, on estimating the β parameters, the covariates pertaining to a given document may be used to obtain the (predicted) easiness of that text (even if it did not appear in sample, or not in that given form).

This is a simple model, and it is worth emphasizing what is being assumed about the data generating process when we interpret its relevant output. First, we assume that the outcomes of the contests are (statistically) independent of one another: that what happens in the kth contest does not affect what happens in the $k+1$th contest. Second, we are making no allowance for variability between snippets which have otherwise identical covariate values. That is, we are not using any kind of random effects for the snippets themselves. This means, equivalently, that the contest results for a given snippet are not modeled as correlated. Third, we make no attempt to include so-called "contest-specific predictors" either in their indirect form—such as effects for (the proclivities of) given human coders—or directly—such as allowing for consequences of the order in which the snippets were presented to the subjects who judged them.

To be clear, the model is sufficiently flexible to allow all these concerns to be addressed—we have simply chosen not to do so. From our perspective, the decision is defensible: our primary interest is in estimating the complexity of documents by predicting (that is, scaling up) from the snippet results. This means we care mostly about the point estimates of the coefficients, not their attendant uncertainty. Since the former is unaffected by our simplification here, that is how we proceed.

1.3.1 Variable Selection via Machine Learning

As noted above, it is not *a priori* obvious which variables should be included in a given model of readability: in this case, the interest is in which features should appear in the linear predictor. To attack this problem, we first fit an *unstructured* Bradley-Terry model, which returns an estimate of an "ability" λ_i (in this case, relative easiness) for each snippet, but makes no use of covariates.[9]

[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, either way.

We then use all our various text characteristics as features to predict these (unstructured) abilities. Specifically, we use a random forests approach (Breiman, 2001), and then we inspect the (relative) 'variable importance' estimates for each covariate. Once those characteristics that matter most are identified, they can be used in the structured model of Equation (2) to obtain the relevant coefficient estimates. We return to the results of that process momentarily.

1.4 A "Bag-of-Snippets" Approach

Finally, as is perhaps obvious by now, ours is a "bag-of-snippets" approach, and is analogous to the "bag-of-words" commonly used in this literature. We assume that the order in which the snippets appear in a document has no effect on how difficult or easy they are to comprehend. Clearly, in some cases this abandoning of context will be an oversimplification. As an example, consider the following two sentences said by one character to another in Tokien's *The Hobbit*:

> There is more in you of good than you know, child of the kindly West. Some courage and some wisdom, blended in measure.

It is not hard to imagine that the second sentence is easier to understand once one has read the first, because the subject of the sentence is implied rather than explicitly stated. In contrast, consider the following two sentence snippet from Joyce's *Finnegans Wake*:

> Begin to forget it. It will remember itself from every sides, with all gestures in each our word.

Here, the first sentence is arguably harder to understand once the second sentence is presented: it is then unclear what *it* might refer to. How often such context issues arise routinely in political texts is an open question, but we ignore them here (or rather, we assume that, on average, their effects are zero).

	AIC	Proportion Correct
FRE	26269.2	0.568
Dale-Chall	26227.9	0.573
FOG	26084.8	0.573
SMOG	26188.2	0.526
Spache	26025.6	0.577
Coleman-Liau	26574.4	0.550

Table 1: Model performance of the standard measures. The overall fit of the Bradley-Terry model using the scores for a given measure is reported in two ways: the Akaike information criterion (AIC) and the Proportion of contest results correctly predicted (where a correctly predicted contest is one in which there is > 0.5 probability that the actual winner would win).

2 Results

We have two main sets of results: in the first part, we compare the standard measures as applied to our political texts. Secondly, and more importantly, we provide a new measure of complexity based on our crowdsourced data and the inferences we draw from our machine learning approach.

2.1 Comparing the Standard Measures

Our setup allows us to compare the performance of the various standard measures in a systematic way. In Table **??** we consider two natural ways to do this. For each of the traditional measures, we fit a Bradley-Terry model which has one predictor: the score for the snippets on a given measure. Thus, the first row refers to a model in which the only covariate is the (difference in the) snippets' Flesch scores (a model we return to below), the second row refers to a model in which the only covariate is the (difference in the) snippets' Dale-Chall scores, and so on. We report the Akaike information criterion for each of these models, along with the proportion of contests correctly predicted by the model. This latter statistic is calculated by generating the relevant λ_is from the linear predictor, using the $\hat{\beta}$ from the model, multiplied by xs for a given snippet. We then calculate the probability that the snippet which actually won a contest would be expected to do so given the estimated parameters—in the sense of Equation 1. If this probability is greater than 0.5, then we declare that a success for the model.

One observations is immediate: the models all perform very similarly, with very little to separate them in terms of either AIC or accuracy. The best performer on our data was the Spache measure, but the FRE is almost exactly as useful and might be preferred on familiarity grounds. We use it in our running comparison for what follows.

2.2 Augmented Bradley Terry Approach

In Appendix A we report the variable importance plots for the random forest models that we ran on the unstructured abilities. Whether we use the Firth (1993) bias reduction technique or not, our conclusions are essentially the same. In particular, the model favors the rarity measure based on the recording the least commonly occurring term in the snippet (relative to 'the' in the Google corpus)—denoted as `google_min_2000`. And it also suggests average sentence length measured in characters (`meanSentenceChars`) is about as important. Given our discussion above, the fact that these variables are useful is unsurprising. In principle, of course, we could stop there (especially given the relatively large distance of the 'top two' from the other variables). In experiments, however, we found that the third most important variable, `pr_noun`—the proportion of words from the text that are nouns—helped model fit. We thus include that one too to form a 'basic' machine learning model.

How does this simple model perform? To assess that, we construct a 'baseline' model which uses the Flesch reading ease (FRE) as its (only) covariate content. We do 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 26269, while the proportion (of contests in the data) correctly predicted (PCP) is 0.568. 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 25912.69, and the proportion correctly predicted rises to 0.583. This is in line with our thinking above: in particular, that models work best when fit 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, PCP is lower).

Studying the model, we note that it doesn't include a measure of word length—yet the success of FRE tells us that almost certainly matters. Looking down the variable importance plots, the first measure of word length to be recommended (i.e. the one 'highest up' in importance terms) is the average number of characters per word (`MeanWordChars`). As an experiment, we added this variable our machine learning model and re-ran the analysis. The results of that process are in the fourth column of Table 2 titled 'Best Model'. Clearly, it now outperforms every other version, with the lowest AIC (25739.93) and the highest PCP (0.587). 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_verb`), but in both cases the fit got worse. This then, is our preferred model for the analysis that follows. Note, in passing, that all the variable 'effects' are expected (and are statistically significant at conventional levels): in particular, *ceteris paribus* texts that contain words which have low (minimum) rarities are easier to understand ('Minimum Google books rarity' is positive), texts that contain longer sentences ('mean Sentence Chars') are harder, and texts with longer words ('mean Word Chars') are also more difficult to comprehend. More nouns ('noun proportion'), on average, also adds to simplicity and this is, in fact, in keeping with earlier work by Flesch (1948) who proposed a 'human interest' index in which a text with more (pro)nouns was generally found to be more compelling than one with fewer.

On what types of data, exactly, does our model do better? Unsurprisingly, given they share core terms, its when two documents are similar other than the proportion of nouns they contain, or the rarity of their words. And, on inspection—i.e. looking at the contests for which our model outperforms the Flesch version to the greatest extent—it's the frequency term that matters. 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, 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.

Thus 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. Put crudely, if researchers think the commonality of terms matters for measuring complexity, our approach is preferred.

It is helpful to be candid about several issues pertaining to our choice here. First, clearly, while we are outperforming the most widely-used measure of readability, our gains are not huge in an absolute sense. Partly this is because the Flesch model is being (re)fit appropriately to the data rather than in its usual 'off-the-shelf' mode. Second, whether or not one uses our *specification*, the general *approach*—of training on relevant data and providing model-based estimates—is surely correct, precisely for the reasons we gave above. Indeed, even if one wanted simply to use the Flesch set up (in terms of its component variables) based on Table 2 we would recommend 'local' data for that purpose.

3 Applications: snippets, *State of the Union* and *Hansard*

We can apply the results of our model in various ways. We outline three obvious approaches before demonstrating how they might be used in practice. First, given Equation (1) and Equation (2), we can obtain a (point) estimate of the probability that any given text i is more difficult than any other text j by calculating

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

	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		-1.78^{*}		
		(0.07)		
Minimum Google books rarity			1310.41^{*}	1332.49^{*}
			(153.27)	(155.85)
mean Sentence Chars			-0.01^{*}	-0.01^{*}
			(0.00)	(0.00)
noun proportion			0.61^{*}	0.63^{*}
			(0.19)	(0.19)
mean Word Chars				-0.31^{*}
				(0.02)
N	19430	19430	19430	19430
AIC	26269.20	25912.69	25917.49	25739.93
PCP	0.568	0.583	0.580	0.587

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 2: Model comparison, post feature-selection. Note that the last column represents our 'optimal' model. 'PCP' is proportion (of contests) correctly predicted by the model.

To see how this works, consider two snippets, one from Eisenhower,

> Here in the District of Columbia, serious attention should be given to the proposal to develop and authorize, through legislation, a system to provide an effective voice in local self-government. While consideration of this proceeds, I recommend an immediate increase of two in the number of District Commissioners to broaden representation of all elements of our local population.

and one from George W. Bush

> 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.

snippet	Min Google rarity	Mean Sent Chars	noun proportion	mean Word Chars
Eisenhower	3.501e-07	158.5	0.23	5.37
Bush	1.40e-08	153.5	0.31	4.72

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

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

$$\lambda_{\text{Eisenhower}} = (1332.49 \times 3.501e-07) + (-0.01 \times 158.5) + (0.63 \times 0.23) + (-0.31 \times 5.37) = -3.10$$

and

$$\lambda_{\text{Bush}} = (1332.49 \times 1.40e-08) + (-0.01 \times 153.5) + (0.63 \times 0.31) + (-0.31 \times 4.72) = -2.80.$$

Following the algebra above, we have

$$\Pr(\text{Eisenhower snippet } \text{ easier than } \text{ Bush snippet}) = \frac{\exp(-3.10)}{\exp(-3.10) + \exp(-2.80)} = 0.425.$$

Of course, these comparisons can be made between *any* two documents—so long as we have covariate values for them—including fifth grade texts, as in Flesch's (1948) original work. In our case, we obtained a set of fifth grade texts from a university education department[10], and estimated the relevant λ as above to be -2.175897. Thus, we the probability that the Eisenhower text is easier than a fifth grade text is estimated to be 0.284; and the probability that the Bush text is easier to follow than the fifth grade works is 0.324. We can place confidence intervals around the point prediction by resampling the sentences in the texts (in the sense of Lowe and Benoit, 2013).

Along with model-based estimates, researchers may also want a quantity analogous to the continuous 0–100 scores from the Flesch (1948) (regression) formula. There are at least two ways to obtain this. First, using Equation (3) denote the $\Pr(i \text{ easier than } j)$ term as p. Then, supposing that we have an appropriate example of a (set of) fifth grade text(s), we can substitute $\exp(\lambda_i)$ for

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

100 (or, indeed, any number preferred) and then rescale $\exp(\lambda_j)$ as $100 \times (\frac{1}{p} - 1)$. Though this preserves the model-based interpretation of the quantity of interest, in practice it tends to return quite low numbers once one is even slightly removed from a fifth grade text. For example, a spotcheck on a document with an FRE of around 84 implies a rescaled score of 35, which 'seems' very low. Again, this is not 'wrong'—it is simply rescaling in a way that preserves the probability structure inherent in the model. But it may well be confusing for end-users, who expect a number approximately commensurate with the grade-level interpretation given by Flesch.

With this problem in mind, an alternative is to simply rescale all the λs (that is, the $\mathbf{X}\beta$s, without applying the exponential function) themselves to be on the relevant interval. For a given data set, a sensible way to proceed is to include a text(s) at the fifth grade level, and one at the post-college level (or whatever minimum and maximum is preferred), and to then scale all resulting λs from 100-0, based on those two end points.[11]

3.1 The Original Snippet Dataset

Experimenting with the continuous measure on the corpus we have performs well in the sense that it returns point estimates on a 0–100 scale that are commensurate (but not identical) to the FRE equivalents. This 'works' because it replaces a logit-style calculation that is not linear in the $\mathbf{X}s$ 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). Clearly the correlation (~ 0.7) is reasonably large and positive. The internal box allows for a more direct comparison of our measure to the (theoretical) minimum and maximum of the FRE: in general, our measure performs similarly. 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 box, particularly to the 'top right' of the plot, our measure tends to score the points differently. Indeed, we assign a considerably higher ('easier')

[11] 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.

rating for the simplest texts.

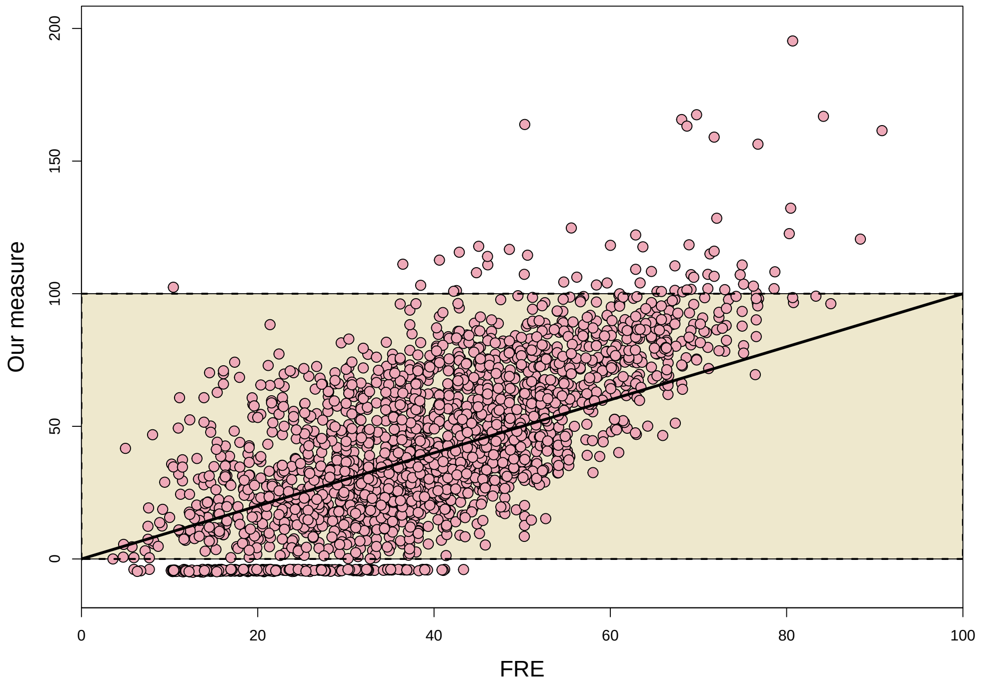

Figure 1: Comparing the 'linear' version of our measure to FRE of the snippets. Correlation is generally high, especially for the theoretical range of the FRE (inner box).

3.2 Reanalyzing the *State of the Union* addresses

Recall that our snippets came from the *State of the Union* time-series, a dataset of considerable interest to academics and journalists. Using our model-based probability measure—here, with a fifth grade text as a baseline for comparison—Figure 2 reports the relevant point estimates and 95% (simulated) confidence intervals (y-axis) plotted against the date of the relevant text. As can clearly be seen, the probability estimates are drifting upwards over time, but generally stay below $\frac{1}{2}$. 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. As can clearly be seen, 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 'dumbing down'—or simplification—may be correct in some aggregate sense if the consider the entire period since the founding of the Republic, but less clear for modern times specifically.

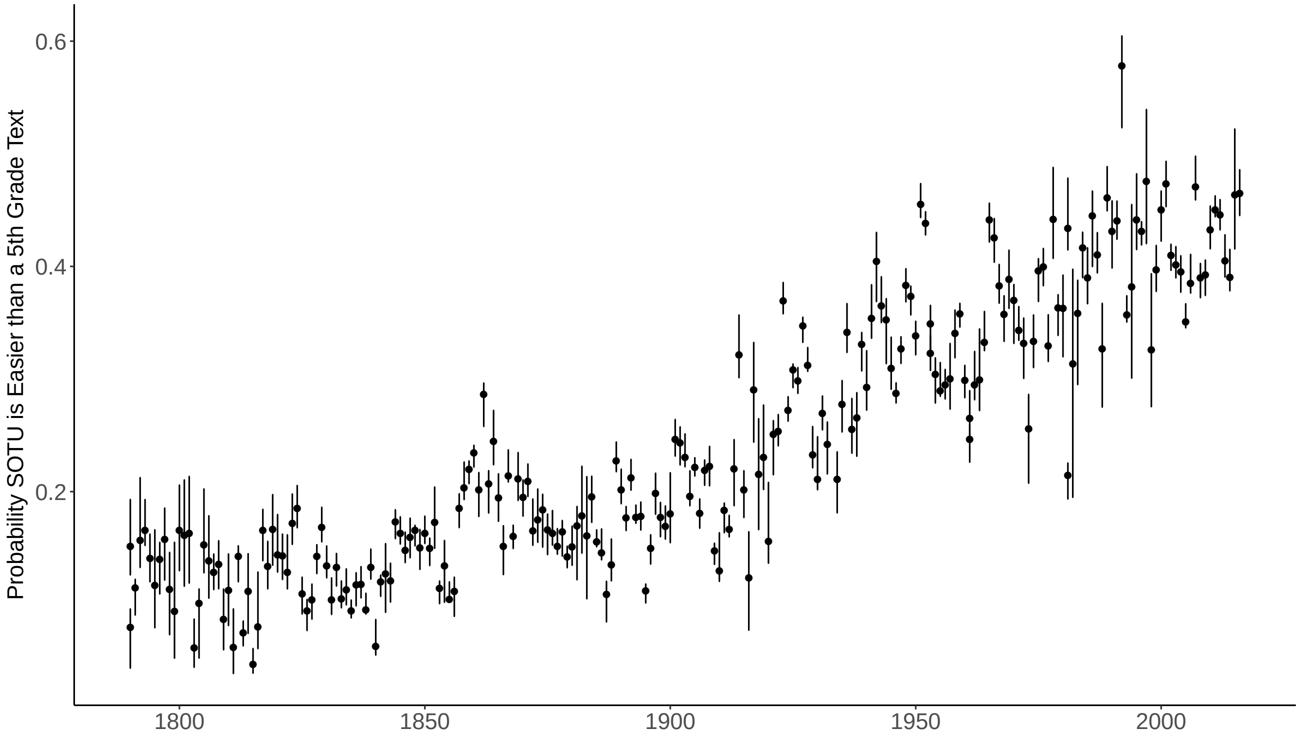

Figure 2:

Of course, since other measures in the literature are not based directly on a statistical model, it is hard to compare our output here with more traditional approaches. Fortunately, the continuous version of our measure does allow a direct comparison, and in 3 (where we label it 'MBE' for model based estimate(s)) we show it plotted against the FRE. Clearly, the conclusions from the measures more-or-less agree for the early history of the Republic: until around 1910, the time series generally overlap. From the first decade of the Twentieth Century, however, our measure accelerates faster and to a higher end point than that of the 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, this makes sense. And, to reiterate, our model is actually trained on appropriate, political data.

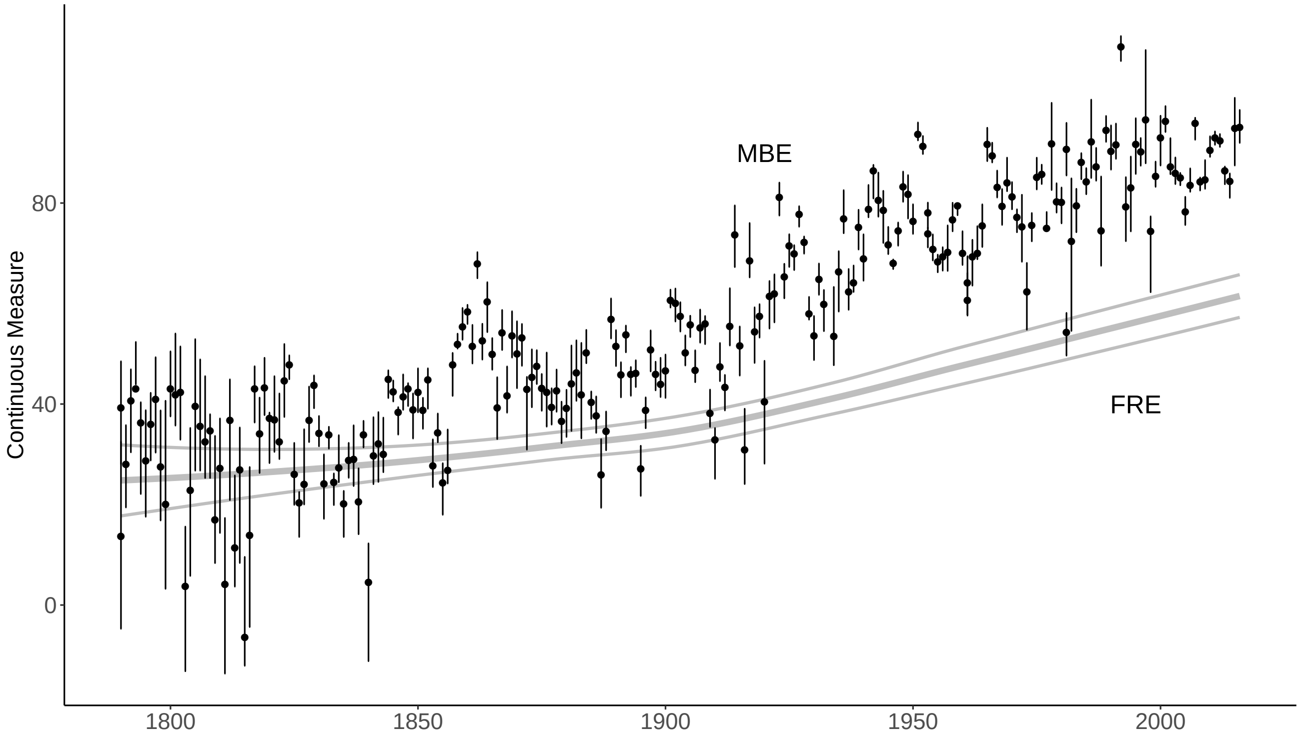

Figure 3: Comparing the ‘linear’, continuous version of our model based estimates (points plus 95% confidence intervals, denoted MBE) to FRE (smooth lines, with outer edges representing 95% confidence intervals) of the State of Union addresses. Note that the time series are initially similar, but depart from one another around the year 1910.

A Random Forest Variable Importance Plots

As noted in text, we ran our random forest model (1000 trees, otherwise standard defaults in the sense of Liaw and Wiener (2002)) for both sets of unstructured estimates—that is, with and without bias-reduction. The results of that process, in terms of the variable importance plots, are given in Figure 4. As usual, variables (on the *y*-axis) with points further right are deemed 'more important' for predicting the outcome (here, the snippet's ability). Notice that the ordering of the variables is similar, regardless of which approach we take (i.e. with or without bias reduction).

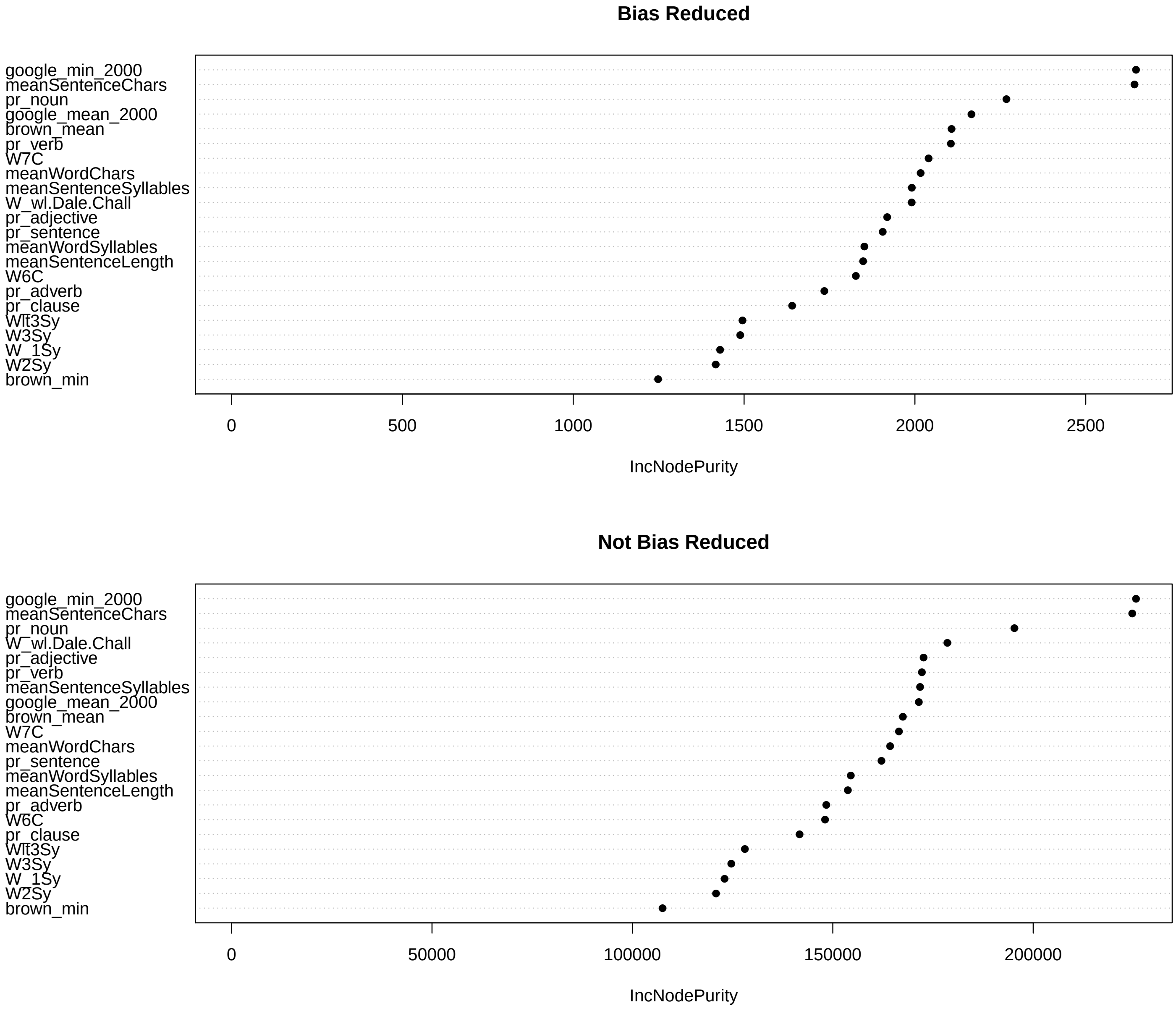

Figure 4: Variable Importance Plots for (unstructured) readability estimates. Note that points further to the right imply more 'important' variables. Top panel is for bias-reduced estimates; bottom panel is for non-bias reduced estimates.

References

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).

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.

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.

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

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

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

Liaw, Andy and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2(3):18–22.
URL: *http://CRAN.R-project.org/doc/Rnews/*

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.

McCullagh, Peter and John Nelder. 1989. *Generalized linear models*. New York: CRC press.

Tränkle, U. and H. Bailer. 1984. “Kreuzvalidierung und Neuberechnung von Lesbarkeitsformeln für die deutsche Sprache.” *Zeitschrift f¨ur Entwicklungspsychologie und Pädagogische Psychologie* 16(3):231–244.

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